Functional classification of protein domain superfamilies for protein function annotation

Sayoni Das

A thesis submitted for the degree of

Doctor of Philosophy

October 2016

Institute of Structural and Molecular Biology

University College London

Declaration

I, Sayoni Das, confirm that the work presented in this thesis is my own except where otherwise stated. Where the thesis is based on work done by myself jointly with others or where information has been derived from other sources, this has been indicated in the thesis.

Sayoni Das

October 2016

Abstract

Proteins are made up of domains that are generally considered to be independent evolutionary and structural units having distinct functional properties. It is now well established that analysis of domains in proteins provides an effective approach to understand protein function using a 'domain grammar'. Towards this end, evolutionarily-related protein domains have been classified into homologous superfamilies in CATH and SCOP databases. An ideal functional subclassification of the domain superfamilies into 'functional families' can not only help in function annotation of uncharacterised sequences but also provide a useful framework for understanding the diversity and evolution of function at the domain level.

This work describes the development of a new protocol (FunFHMMer) for identifying functional families in CATH superfamilies that makes use of sequence patterns only and hence, is unaffected by the incompleteness of function annotations, annotation biases or misannotations existing in the databases. The resulting family classification was validated using known functional information and was found to generate more functionally coherent families than other domain-based protein resources. A protein function prediction pipeline was developed exploiting the functional annotations provided by the domain families which was validated by a database rollback benchmark set of proteins and an independent assessment by CAFA 2.

The functional classification was found to capture the functional diversity of superfamilies well in terms of sequence, structure and the protein-context. This aided studies on evolution of protein domain function both at the superfamily level and in specific proteins of interest. The conserved positions in the functional family alignments were found to be enriched in catalytic site residues and ligand-binding site residues which led to the development of a functional site prediction tool. Lastly, the function prediction tools were assessed for annotation of moonlighting functions of proteins and a classification of moonlighting proteins was proposed based on their structure-function relationships.

Acknowledgements

The last four years of my life as a PhD student has been a remarkable experience. It would not have been possible for me to finish my doctoral research without the support and guidance that I received from many people.

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Christine Orengo for her invaluable guidance, patience and encouragement throughout the period of my doctoral research and for giving me the opportunity to work in such a friendly and prestigious group. I am very grateful to her for being an excellent teacher and mentor.

I would like to acknowledge UCL for the UCL Overseas Research Scholarship. I would like to thank my Thesis committee members, Dr. Andrew Martin and Prof. Snezana Djordjevic for their for their valuable insights and support during my thesis committee meetings. I would also like to thank Ishita Khan and Prof. Daisuke Kihara for helpful discussions on moonlighting proteins.

I would like to thank all the past and present members of the Orengo group for always being kind and helpful and for making my stay in the group such a pleasant experience. In particular, thanks to Dr. Ian Sillitoe and Dr. Natalie Dawson for their help and advice throughout my research. Special thanks to Dr. David Lee and Dr. Jon Lewis for their help and support during CAFA 2. Thanks to Dr. Tony Lewis, Dr. Romain Studer and Dr. Paul Ashford for their help and advice with any computational or programming questions. Thanks are also due to Dr. Rob Rentzsch for his help when I first started with my research work. Special thanks to Francesco Carbone, Ivana Pilizota, Su Datt Lam, Tom Northy, Sonja Lehtinen, Saba Ferdous, Avneet Saini, Millie Pang for many amazing times. I would like also like to thank Jesse from the Biosciences IT and David Gregory and Tristan Clark from the CS Cluster Support for their timely support and help throughout my research work. Finally, a big thank you to everyone in the Room 627 and the rest of Institute of Structural and Molecular Biology for always being friendly and helpful.

My fellow PhD student and best friend Priscilla has also played a big role in

making my stay in London an enjoyable experience. I am ever so grateful to her for being there for me and for always being a source of hope and positivity. Special thanks to Arghya for being patient, understanding, encouraging and for his efforts for teaching me to be more pragmatic. Thanks to Ambalika and Kalpana for their love, understanding and support throughout my research work. I would like to thank Prateek, Pavi, Challenger, Abhishek, Anjul and Tanvi for many a wonderful times in the last four years. Also, thanks to all my other friends for their best wishes.

I would like to thank my cousin, aunt and uncle for their love and for making me feel at home in London. Last but not the least, heartfelt thanks to my parents, my brother and my sister-in-law for their unwavering love and support at the most difficult times. Special thanks to the little bundle of joy in our family, Tara, who is growing up at an alarming rate and makes me smile whenever I see her latest picture or video. I am grateful to them for having faith in me. Finally, I would like to dedicate this thesis to my family.

Contents

List of Figures

List of Tables

List of Abbreviations

ADDA Automated Domain Delineation Algorithm

BDT Binding site distance test

BLAST Basic Local Alignment Search Tool

BLOSUM Blocks Substitution Matrices

BPO Biological Process Ontology

CASP Critical Assessment of Methods of Protein Structure Prediction

CAFA Critical Assessment of Protein Function Annotation

CATH Class, Architecture, Topology, Homologous Superfamily

CCO Cellular Compartment Ontology

CDD Conserved Domain Database

COG Clusters of Orthologous Groups

CSA Catalytic Site Atlas

DAG Directed Acyclic Graph

DFX Domain Family Exploration

DOPS Diversity of Position Score

EC Enzyme Commission

ECOD Evolutionary Classification of protein Domains

E-value Expectation value

ET Evolutionary Trace

FC Function Coherence index

FD Functional determitant

FunFam Functional family

GeMMA Genome Modelling and Model Annotation

GO Gene Ontology

HAD Haloalkanoic acid dehalogenase

HMM Hidden Markov Model

IBIS Inferred Biological Interactions Server

IC Information content

IEA Inferred from Electronic Annotation

KEGG Kyoto Encyclopedia of Genes and Genomes

LIG Ligand-binding

LK Limited-knowledge

MACiE Mechanism, Annotation, and Classification in Enzymes

MCC Matthews Correlation Coefficient

MDA Multi-domain Architecture

MFO Molecular Function Ontology

MSA Multiple sequence alignment

NCBI National Center for Biotechnology Information

NK No-knowledge

PAM Point Accepted Mutation

PDB Protein Data Bank

PP Pyrophosphate

PPI Protein-protein interaction

PSI-BLAST Position Specific Iterated-BLAST

PSSM Position-Specific Scoring Matrix

PYR Pyrimidine

RMSD Root Mean Squared Deviations

SCI-PHY Subfamily Classification In Phylogenomics

SCOP Structural Classification Of Proteins

SDP Specificity Determining Position

SFLD Structure Function Linkage Database

SSAP Sequential Structure Alignment Program

SSG Structurally-Similar Group

SVM Support Vector Machine

TEED Thiamine-diphosphate dependent Enzyme Engineering Database

TPP Thiamine pyrophosphate

TrEMBL Translation of coding sequences from the EMBL database

UniProtKB UniProt Knowledgebase

List of Publications

Lee, D., **Das, S.**, Dawson, N. L., Dobrijevic, D., Ward, J. and Orengo, C. (2016). Novel computational protocols for functionally classifying and characterising serine beta-lactamases, *PLoS Computational Biology*, **12**.6, e1004926.

Lam, S. D., Dawson, N. L., **Das, S.**, Sillitoe, I., Ashford, P., Lee, D., Lehtinen, S., Orengo, C. A. and Lees, J. G. (2016). Gene3D: expanding the utility of domain assignments., *Nucleic Acids Research*, **44**.D1, D404–9.

Das, S. and Orengo, C. A. (2016). Protein function annotation using protein domain family resources, *Methods*, **93**, 24–34.

Das, S., Dawson, N. L. and Orengo, C. A. (2015). Diversity in protein domain superfamilies, *Current Opinion in Genetics & Development*, **35**, 40–49.

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G. and Orengo, C. A. (2015). Functional classification of CATH superfamilies: a domain-based approach for protein function annotation, *Bioinformatics*, **31**.21, 3460–3467.

Das, S., Sillitoe, I., Lee, D., Lees, J. G., Dawson, N. L., Ward, J. and Orengo, C. A. (2015). CATH FunFHMMer web server: protein functional annotations using functional family assignments, *Nucleic Acids Research*, gkv488.

Sillitoe, I., Lewis, T. E., Cuff, A., **Das, S.**, Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G. et al. (2015). CATH: comprehensive structural and functional annotations for genome sequences, *Nucleic Acids Research*, **43**.D1, D376–D381.

Lees, J. G., Lee, D., Studer, R. A., Dawson, N. L., Sillitoe, I., **Das, S.**, Yeats, C., Dessailly, B. H., Rentzsch, R. and Orengo, C. A. (2014). Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis, *Nucleic Acids Research*, **42**.D1, D240–D245.

Chapter 1

Introduction

Proteins are important biomolecules which carry out the biological functions required for proper functioning of the cell. Their function can range from biological catalysts or enzymes (e.g. trypsin, pepsin), structural elements (e.g. keratin, collagen), transport molecules (e.g. haemoglobin), storage molecules (e.g. ferritin) inside the cell through to other important roles in biological processes such as cell-signalling (e.g. kinases, phosphatases), immune response and cellular metabolism amongst others. Therefore, characterization of the functions of all proteins is key to having a better understanding of the cell at the molecular level. Moreover, characterisation of protein function has huge pharmaceutical and biotechnological implications. The availability of large number of protein sequences and structures and the use of computational approaches for protein function annotation has been a significant step towards this. However, the complex evolution of function in proteins and their increasing diversity presents new challenges to existing protein function annotation methods. Large-scale functional classification of protein resources can not only improve the reliability of function prediction for uncharacterised sequences, but can also help in understanding how function is modulated during evolution by sequence and structural changes.

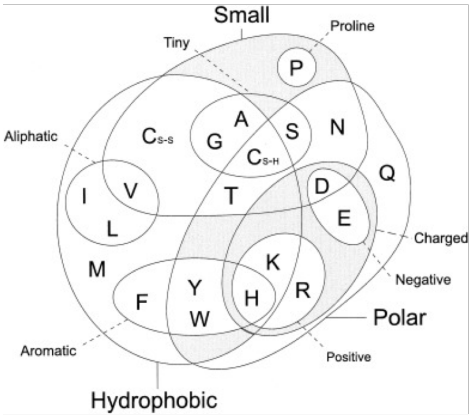
In this thesis, work is presented which describes a new approach for functional classification of protein domain superfamilies in the CATH-Gene3D resource. The functional classification captures the functional diversity of protein domains which is exploited to provide functional annotations for uncharacterised proteins using a domain-based approach and to understand the mechanisms of functional divergence in domain superfamilies during evolution.

This chapter describes general bioinformatics concepts, methods and resources that are relevant to the work presented in this thesis, followed by an outline of the following chapters.

1.1 Protein function

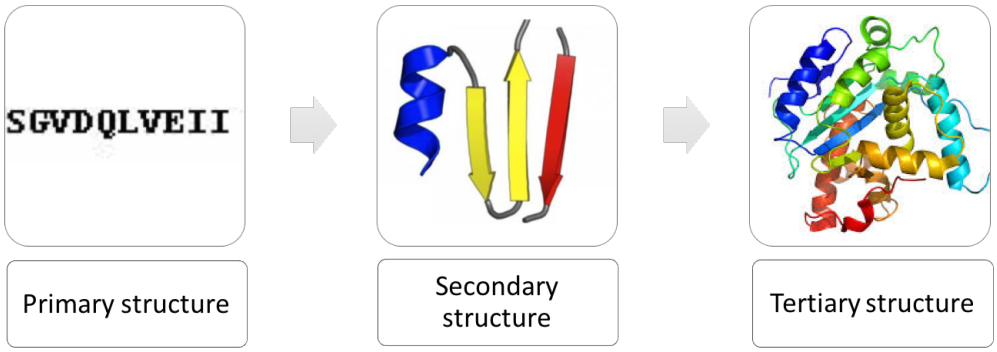
Proteins are formed of one or more linear polymers built from a library of 20 different amino acids linked together by peptide bonds (polypeptide chain). The amino acid sequence of a protein forms the primary structure of the protein and is determined by the nucleotide sequence of the gene encoding it. The physico-chemical characteristics of different amino acid side chains have important implications in the structure and function of proteins (Figure 1.1a). The polypeptide chain forms regular secondary structures such as α-helices and β-sheets as a result of hydrogen bonding interactions between the main chain peptide groups interspersed with regions of irregular structures such as loop or coil regions. The linear secondary structure of the proteins, then folds into the native three-dimensional (3D) conformation known as the tertiary structure that allows proteins to interact with other proteins or molecules and perform their function (Figure 1.1b). Furthermore, for proteins constituting more than one polypeptide chain, the relative arrangement of two or more polypeptide chains in the protein forms the quaternary structure.

Proteins are generally composed of one or more building blocks called domains that are considered to be the structural, evolutionary and functional units of proteins. The definition of a domain can however vary slightly between databases. For example, in the CATH (Orengo *et al.*, 1997) database, domains are considered to be distinct, compact units of protein structure (Figure 1.1c), in the SCOP (Murzin *et al.*, 1995) database, domains are considered to be independent evolutionary units. Additionally, as a functional unit, domains are considered to have an independent function, however, sometimes they contribute to the function of a multi-domain protein in combination with other domains (Vogel *et al.*, 2004). In the present work, protein domains are considered to be continuous or discontinuous regions of sequence or structure forming compact structural units that are conserved between related proteins and all domain definitions used in this study are taken from the CATH database.

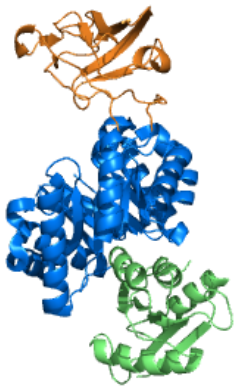


Amino Acid	3-Letters	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

(a)



(b)



(c)

Figure 1.1: (a) Taylor's Venn diagram (Taylor, 1986) of amino acid properties illustrating the physico-chemical similarities and differences of the 20 different amino acids (listed on the right). Taken from Valdar (2002). (b) The primary, secondary and tertiary levels of protein structure. (c) Structure of *E. coli* pyruvate kinase, a protein composed of three structural domains (shown in orange, blue and green).

Protein domains often combine with other domains in a mosaic manner as a result of gene duplication, combination and fusion, in a process known as 'domain shuffling', giving rise to multi-domain proteins with new or modified functions (Chothia *et al.*, 2003) (see Section 1.3.1.3 for details). Within a multiple-domain protein, different domains tend to have different functional roles which in combination make up the overall function of the protein (Bashton and Chothia, 2007). Since multi-domain proteins expand the functional repertoire, this further complicates protein sequence-structure-function relationships. At least two thirds of eukaryotic and more than a half of prokaryotic proteins are composed of multiple domains (Han *et al.*, 2007).

1.1.1 Multi-faceted function of proteins

The function of a protein is context-based and can be studied from different aspects, ranging from biochemical activity to the role of the protein in pathways, cells, tissues and organisms. The phrase 'protein function' is very ambiguous as the functional role of a protein can be described in many different contexts. It can be explained in terms of: (i) the molecular function of the protein, (ii) its role in a biological pathway and (iii) its cellular location. Natural language annotations in databases and literature were found to be too vague and unspecific to accurately describe the function of proteins. This has led to the need and subsequent development of several resources for organized protein annotation vocabularies.

1.1.2 Function annotation resources

Various protein annotation schemes have been developed like the Enzyme Commission (EC) number (Bairoch, 1994), KEGG (Kanehisa *et al.*, 2015), FUNCAT (Functional Catalogue) (Ruepp *et al.*, 2004) and Gene Ontology (GO) (Ashburner *et al.*, 2000). While KEGG and FUNCAT have traditionally been more focused on providing annotations of biological processes, the EC number system and the Gene Ontology scheme are the most widely used protein function annota-

tion resources. Complementary to these annotation resources, the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004) provides information on catalytic residues for enzymes of known structure, the NCBI Inferred Biomolecular Interactions Server (IBIS) reports experimentally determined interaction binding interfaces and the MACiE (Mechanism, Annotation and Classification in Enzymes) (Holliday *et al.*, 2007) database provides detailed information on enzyme reaction mechanisms.

1.1.2.1 Enzyme Commission number

The Enzyme Commission (EC) number system was one of the first sources of protein annotations (Webb, 1992; Bairoch, 1994) and is now well established. It is a numerical classification system for enzymes representing the catalytic function using a hierarchical classification of four levels to represent the catalytic reaction that it carries out. For example, in EC number 1.2.1.1, the first number (EC 1.-.-.-) refers to the enzyme class - oxidoreductases in this case, the second (EC 1.2.-.-) refers to the type of bond that is acted on i.e. the aldehyde or oxo group of donors, the third (EC 1.2.1.-) refers to details like electron acceptor which is NAD^+ or $NADP^+$ in this case and the fourth number (EC 1.2.1.1) refers to the substrate formaldehyde dehydrogenase. However, the EC number system does not provide any classification for non-enzymatic proteins. Additionally, it is not sufficient to either describe all roles of a protein within a cell or the diverse interactions of a protein inside the cell.

1.1.2.2 Gene Ontology

The Gene Ontology (GO) (Ashburner *et al.*, 2000) is the largest and the most-widely used resource of protein annotations and is managed by the Gene Ontology Consortium. The sources of the GO annotations can be literature references, database references or computational evidences which are indicated by the GO evidence codes (see Table 1.1) associated with each annotation.

Table 1.1: Evidence codes associated with GO annotations.

GO evidence codes	Definition
EXP	Inferred from Experiment
IDA	Inferred from Direct Assay
IPI	Inferred from Physical Interaction
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IEP	Inferred from Expression Pattern
TAS	Traceable Author Statement
IC	Inferred by Curator
IEA	Inferred from Electronic Annotation
ISS	Inferred from Sequence or Structural Similarity
ND	No biological Data available

The GO annotations can be used to assign functional terms to both enzymes and non-enzymes from three structured, non-overlapping ontologies or vocabularies in a species-independent manner:

(i) **Molecular Function Ontology (MFO)** describes the biochemical activity of the protein at the molecular level;

(ii) **Biological Process Ontology (BPO)** describes the cellular processes and pathways in which the protein is involved;

(iii) **Cellular Component Ontology (CCO)** describes the compartments(s) of the cell in which the protein performs its action.

Each individual GO ontology has a tree-like structure of a hierarchical Directed Acyclic Graph (DAG), where the GO functional terms (nodes) are organized bottom-up by child-parent relationships (forming the edges of the DAG) like ‘is a’,‘part of’, ‘has part’ or ‘regulates’ (Figure 1.2). In each ontology, any path from a particular GO term towards the root becomes more general as they become subsumed by the parent terms. Using GO, a gene or protein is annotated by associating the most specific set of GO terms which describes its function accurately. Furthermore, when a gene or protein is associated with a

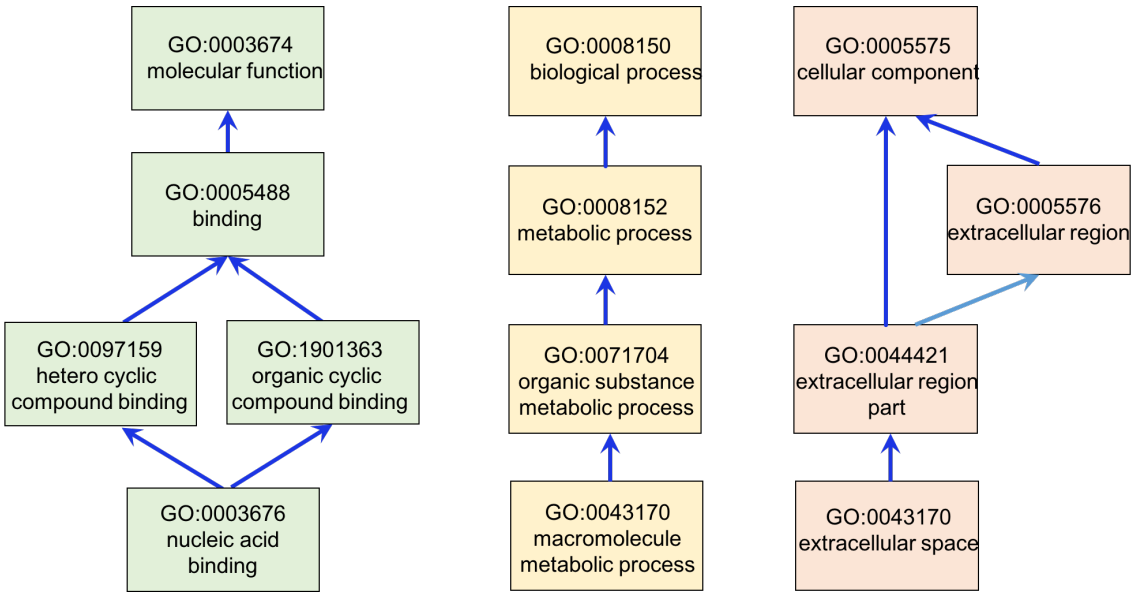


Figure 1.2: Three categories of Gene Ontology (GO): (i)Molecular Function Ontology (MFO), (ii) Cellular Component Ontology (CCO) and (iii) Biological Process Ontology (BPO). The dark blue lines show 'is a' relationships while the light blue lines show 'part of' relationships.

particular term, following its ontology structure, it also gets implicitly associated with all its ancestral terms up to the root of the ontology (propagation of GO annotations). For example, if a protein is annotated with the term GO:0003676 (*nucleic acid binding*) from the Molecular Function Ontology (Figure: 1.2(i)), it also inherits all its parental MFO terms: GO:0097159 (*organic compound binding*), GO:1901363 (*heterocyclic compound binding*), GO:0005488 (*binding*) and GO:0003674 (*molecular function*).

The Gene Ontology is a powerful tool for protein annotation analysis which offers many advantages over other sources (Rentzsch and Orengo, 2009). The main advantage of GO annotations is that it allows quantitative comparison of functional similarity between proteins (or genes) using various semantic similarity measures. Semantic similarity between GO terms can be quantitated using either information content-based or graph-based methods. The information content-based methods (Jiang and Conrath, 1997; Resnik and Yarowsky, 1999; Lin, 1998; Schlicker *et al.*, 2006) used to compute semantic similarity depend on the annotation frequencies of the common ancestor terms of GO terms. Informa-

tion content (IC) of a GO term is defined as the the negative log probability of the term occurring in the ontologies (Equation 1.1). In other words, a rarely used GO term in the ontologies will contain a greater amount of information compared to frequently used terms.

The information content (IC) of a GO term t_i in an ontology is defined as:

$$IC(t_i) = -log(p(t_i)) \tag{1.1}$$

where $p(t_i)$ is the probability of the GO term t_i occurring in the ontology.

$p(t_i)$ is estimated using the relative frequency of occurrence of the GO term t_i or its children terms in the ontology:

$$p(t_i) = freq(t_i)/N \tag{1.2}$$

where *N* refers to the total number of annotations (with any GO term) within the ontology and $freq(t_i)$ is the frequency of occurrence of the GO term t_i or its children terms (i.e. the set of all terms for which t_i is a parent term) in the ontology such that

$$freq(t_i) = |n(t_i)| + \sum_{c \in children(t_i)} |n(c)| \tag{1.3}$$

where *n* is the number of annotations with a particular GO term (Mistry and Pavlidis, 2008).

In contrast, the graph-based method by Wang *et al.* (2007), makes use of the topology of the GO graph structure to measure similarity between the terms. The Gene Ontology can also be used as a database to look up genes or proteins having similar functions and can also be used to infer the function of unannotated genes or protein-protein interactions.

Recently, significant annotation biases have been identified in GO annotations (Schnoes *et al.*, 2013). These biases have arisen from the recent increase in use of high-throughput experiments in functional annotation of whole genomes

(functional genomics) which contribute substantially (up to 25%) to experimental protein annotations. High-throughput experiments provide annotations that have lower information content (Equation 1.1) compared to low or moderate-throughput experiments and are biased towards providing a limited number of function annotations. Since biased experimental annotations can only provide a partial picture of the function of a protein, it not only affects our understanding of the protein function space but also affects the relationship between protein function prediction methods and the predicted protein function (Schnoes *et al.*, 2013). Hence, it is essential for both developers of automated function prediction algorithms and biologists who use computational function annotations to guide their experiments, to be aware of the existence of such experimental biases.

1.1.3 Widening function annotation gap

In this post-genomic era, due to the increasing genome-sequencing initiatives worldwide and cheaper associated costs, a huge amount of nucleic acid and protein sequences are accumulating in our databases (see Figure 1.3). As of January 2016, the protein database UniProt Knowledgebase (UniProtKB) (UniProt-Consortium, 2015), comprising the unreviewed UniProtKB/TrEMBL and the manually-curated UniProtKB/Swiss-Prot, contains more than 55 million protein sequences. In contrast, only $\sim$ 1% of the sequences have been experimentally characterised. Additionally, the protein structure data has also expanded to a great extent in the last decade as a result of structural genomics initiatives. The Protein Data Bank (PDB) (Rose *et al.*, 2015) which is the global repository of protein structures currently holds more than 106,000 structures and has grown by almost 20% in the last two years alone. Since the experimental function annotation of such large amounts of sequence and structural data is not feasible, the function annotation gap will continue to widen. In order to bridge this gap, bioinformatics methods and protocols for function annotation of proteins have become essential.

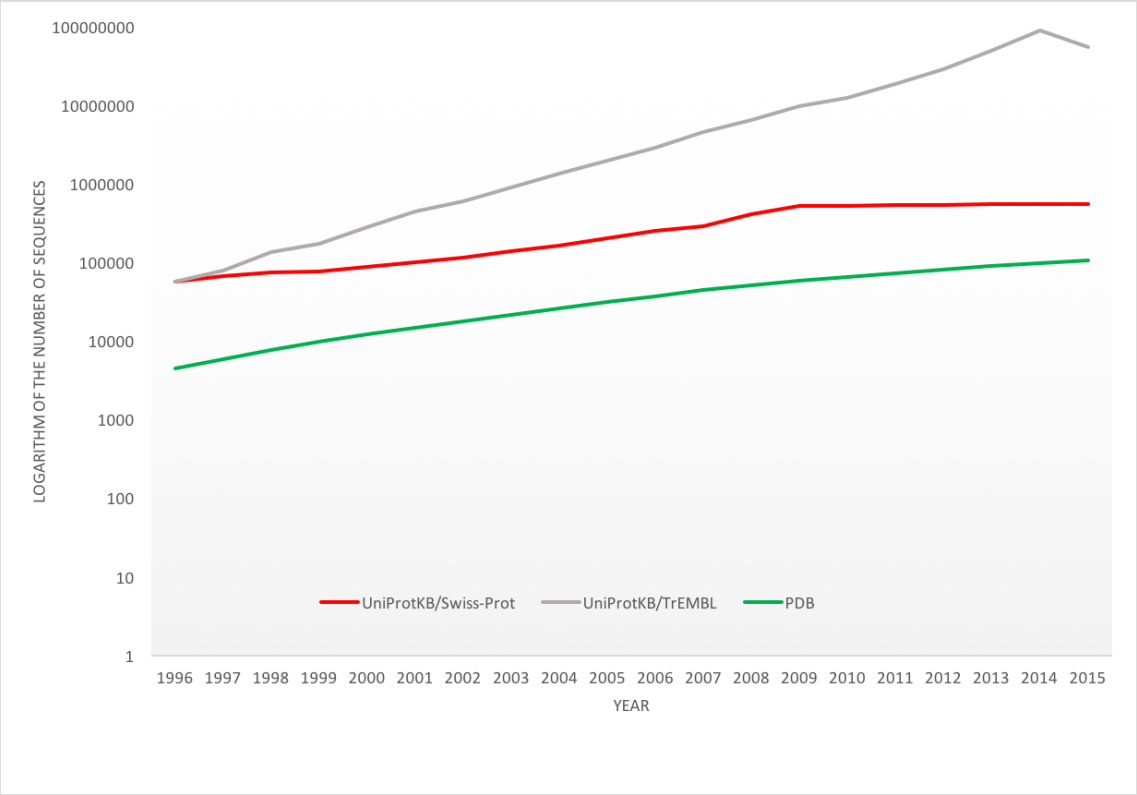


Figure 1.3: The yearly growth of UniProtKB/TrEMBL, UniProtKB/Swiss-Prot and PDB from 1996 to 2015 in the logarithmic scale. A large number of sequences corresponding to redundant proteomes were removed from UniProtKB/TrEMBL in 2015 which resulted in the sudden depletion of sequences in the TrEMBL database (shown as grey line) in 2015.

1.2 Bioinformatics methods and protocols

1.2.1 Protein sequence analysis

Protein sequence analysis is useful for discovering structural, functional and evolutionary information about a protein. Evolutionarily-related sequences i.e. sequences related by divergence from a common ancestor are defined as being homologous. As homologous proteins tend to have similar structures and functions, the inference of sequence homology of an uncharacterised protein sequence to a sequence of characterised function is of utmost importance to protein structure and function prediction.

Homologous sequences can be either orthologous or paralogous. Orthologs are homologs which arise as a result of a speciation event and they generally tend

to retain the ancestral function in different species. In contrast, paralogs arise from a gene duplication event within a genome in which one gene tends to retain the original function and the other, free from selective pressure, can rapidly evolve a new function. Although, orthologs are generally more functionally similar than paralogs, that may not always be true as functional divergence between orthologs has also been reported by various studies (Studer and Robinson-Rechavi, 2009; Altenhoff *et al.*, 2012).

1.2.1.1 Sequence alignments

Sequence alignment is the procedure of comparing two (pairwise sequence alignment) or more (multiple sequence alignment) sequences for identifying regions of similarity in the sequences by searching for a series of characters (amino acids for protein sequences) that are in the same order. As the similarity between sequences can have structural, functional or evolutionary implications, it is essential to obtain a reliable alignment of sequences to discover these relationships. In this work, only protein sequence alignments will be discussed.

In a sequence alignment, the aligned sequences are represented in individual rows, in which gaps are inserted such that identical or similar residues of the sequences are aligned in successive columns. For homologous sequences, gaps in an alignment can be interpreted as insertions or deletions (indels) and mismatched residues can be interpreted as point mutations introduced in one or more lineages in the time since they diverged from one another. Hence, as more distantly related sequences are aligned, more gaps and mutations can be expected in the alignment.

There can be two types of sequence alignments - global or local (Figure 1.4). In global sequence alignment, an end-to-end alignment of sequences are generated which is useful for aligning similar sequences of approximately the same length. In local sequence alignment, stretches of sequences sharing the highest similarity are aligned resulting in one or more sub-alignments within the align-

ment. This type of alignment is useful for aligning sequences of different lengths or sequences sharing a conserved region or domain.

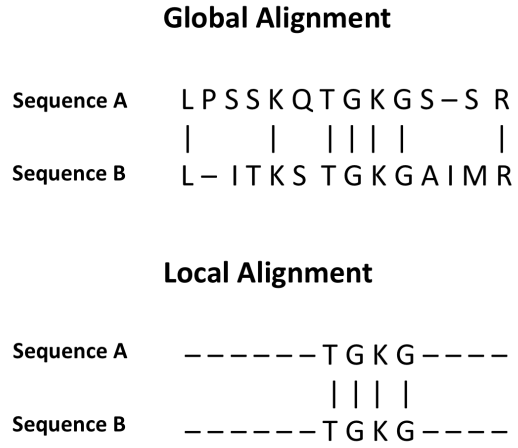


Figure 1.4: Example of global and local pairwise sequence alignments.

Amino acid substitution matrices, such as the Dayhoff Point Accepted Mutation matrix 250 (PAM250) (Dayhoff and Schwartz, 1978) or BLOSUM substitution matrix 62 (BLOSUM62) (Henikoff and Henikoff, 1992) is used to score the matches and mismatches in the alignments. These matrices consist of all-against-all amino acid scores that reflects how often one amino acid would have been mutated to the other over a given evolutionary timescale from an alignment of related protein sequences. The choice of a suitable scoring matrix greatly affects the performance of the alignment method. The PAM matrices lists the likelihood of change from one amino acid to another in homologous sequences during evolution that were estimated from an analysis of 71 groups of closely related protein sequences sharing at least 85% similarity. On the other hand, the scores in the BLOSUM matrices are based on observed amino acid substitutions in a large set of conserved amino acid patterns (known as blocks) derived from $>$ 500 protein families. The PAM matrices are designed to track the evolutionary origins of proteins, whereas the BLOSUM model is designed to find their conserved domains. In addition to the Dayhoff PAM and BLOSUM matrices, a number of other amino acid substitution matrices are widely used for producing

protein sequence alignments such as JTT (Jones *et al.*, 1992) and WAG (Whelan and Goldman, 2001) matrices.

Pairwise sequence alignments

The simplest sequence comparison method is the dot matrix analysis that uses a two-dimensional matrix to visualize the similarity between two protein sequences, in which the residues of the two protein sequences are mapped along the horizontal and vertical axes respectively. For a simple visualisation, individual cells in the matrix can be shaded black if residues are identical, such that matching sequence segments appear as runs of diagonal lines across the matrix. The aim of the dotplot analysis is to estimate a single path through the matrix which has the most biological significance, i.e. the path which aligns the most identical residues or residues which are expected to be tolerated as mutations. Methods based on dynamic programming approaches (Needleman and Wunsch, 1970; Smith and Waterman, 1981) were subsequently devised for finding this optimal path. This involved conversion of the dot plot to a score matrix or path matrix, in which cells are scored according to the similarity of the residue pairs associated with them. The optimal path is then calculated as the one having the highest score.

The Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) and the Smith-Waterman algorithm (Smith and Waterman, 1981) are used for generating global and local alignments respectively. The dynamic programming methods for pairwise sequence alignment first populates a two dimensional matrix with scores according to the identities or similarities of residues associated with each cell by following a scoring scheme for matches, mismatches, and gaps. The scores in the matrix are then accumulated from the bottom right corner of the matrix to the top left corner of the matrix. Once the matrix has been accumulated, the highest scoring path through the matrix is traced back from top left to bottom right which reflects the optimal pair-wise alignment.

For two sequences $a = a_1, a_2...a_n$ and $b = b_1, b_2...b_n$, the score H_{ij} at position *i*

in sequence *a* and position *j* in sequence *b* can be for global and local alignments are shown below $-$

Global alignment (Needleman and Wunsch, 1970):

$$\begin{aligned} H_{ij} \quad = max\{ \quad & H_{i-1,j-1} + s(a_i, b_j), \\ & \max_{x \geq 1} (H_{i-x,j} - w_x), \\ & \max_{y \geq 1} (H_{i,j-y} - w_y) \ \} \end{aligned} \tag{1.4}$$

Local alignment (Smith and Waterman, 1981):

$$\begin{aligned} H_{ij} \quad = max\{ \quad & H_{i-1,j-1} + s(a_i, b_j), \\ & \max_{x \geq 1} (H_{i-x,j} - w_x), \\ & \max_{y \geq 1} (H_{i,j-y} - w_y), \\ & 0 \ \} \end{aligned} \tag{1.5}$$

where $s(a_i, b_j)$ is the score for aligning the characters at positions *i* and *j*, w_x is the penalty for a gap of length *x* in sequence *a*, and w_y is the penalty for a gap of length *y* in sequence *b*.

Multiple sequence alignments (MSAs)

Finding an optimal alignment for more than three sequences using dynamic programming becomes very computationally expensive. Therefore, to align a large number of sequences in a reasonable amount of time, heuristic algorithms have been used. Various methods using progressive and iterative heuristics are available for obtaining reliable multiple sequence alignments (MSAs).

Progressive alignment methods build an MSA by gradually combining a series of pairwise alignments starting from the most similar pair to the most distantly

related one. The sequence relationships are first represented as a tree known as the guide tree, followed by addition of sequences sequentially to the growing MSA according to the branching order of the guide tree. As a result, any errors made in the earlier stages of building the MSA are also propagated to the final MSA. Use of accurate guide trees is critical to the performance of the method and the method of generating them varies with the alignment methods. The most popular progressive multiple sequence alignment tools is Clustal Omega (Sievers *et al.*, 2011). Alternatively, iterative alignment methods realign subgroups of sequences followed by alignment of these subgroups into a global alignment of all of the sequences. The selection of the sub-sequences can be based on a tree similar to progressive methods or on a random selection. Examples of a widely-used iterative methods is MUSCLE (Edgar, 2004).

The MAFFT (Katoh *et al.*, 2002) suite of MSA programs that combines both the progressive and iterative approaches, and is both faster and more accurate than many alignment methods. The MAFFT programs are based on the Fast Fourier transform (FFT) in which a protein sequence is converted to a sequence composed of volume and polarity values of each amino acid residue which allows rapid detection of homologous regions. Different progressive (FFT-NS-1, FFT-NS-2) and iterative refinement methods (FFT-NS-i, L-INS-i, E-INS-i) have been implemented in a range of MAFFT programs ranging from highly accurate for $<$200 sequences (L-INS-i) to very fast (FFT-NS-2) for $>$2000 sequences. Several independent benchmarks (Nuin *et al.*, 2006; Thompson *et al.*, 2011) have shown MAFFT (L-INS-i) outperforming other MSA methods such as ProbCons (Do *et al.*, 2005), MUSCLE (Edgar, 2004), ClustalW (Thompson *et al.*, 2002) and T-Coffee (Notredame *et al.*, 2000).

1.2.1.2 Identification of conserved residues

The neutral theory of molecular evolution suggests that most of the sites in a protein undergo random mutations which do not affect its function while only a few

sites are under stringent evolutionary constraints reflected by absence of substitutions or substitution of biochemically similar amino acids. Hence, the degree of conservation of a position in a multiple sequence alignment indicates the structural or functional importance of that position (see Figure 1.5). Positions implicated with a structural role may be important for protein folding and stability while positions under functional constraints may be catalytic or involved in ligand binding, protein-protein interactions and protein-DNA binding. Many automated methods have been developed to quantify residue conservation and identify conserved positions based on residue frequency scores (Wu and Kabat, 1970; Jores *et al.*, 1990), entropy-based scores (Shannon, 2001), stereochemical property-based scores (Taylor, 1986), mutation data-based scores (Karlin and Brocchieri, 1996) and weighted scores (Valdar, 2002).

Scorecons (Valdar, 2002) quantifies the conservation of each residue position in a protein multiple sequence alignment. Each position in an alignment is assigned a conservation score between 0 and 1, where a score of 0 indicates no conservation at that position and a score of 1 indicates that the residue at that position is completely conserved. Scorecons calculates a weighted sum of pairs score i.e. the sum of all possible pairwise similarities between residues in an aligned column, using amino acid substitution probabilities from the Pairwise Exchange Table (PET91) (Jones *et al.*, 1992), a Dayhoff-like mutation data matrix. The Scorecons score for position *i* in an alignment is defined as:

$$Scorecons(i) = \frac{\sum_{j}^{N} \sum_{k>j}^{N} W_j W_k Mut(s_j(i), s_k(i))}{\sum_{j}^{N} \sum_{k>j}^{N} W_j W_k} \tag{1.6}$$

where *N* is the number of sequences in the alignment, $s_j(i)$ and $s_k(i)$ on the basis of: (i) weighted symbol diversity calculated by an entropy-related measurement (Shannon's entropy), (ii) stereochemistry diversity calculated using amino acid similarity information from a substitution matrix and (iii) the cost of gaps in the alignment.

Scorecons also provides a Diversity of Position Score (DOPS) between 0 and

100 that considers the number of different conservation scores in an alignment and the relative frequency of each score, reflecting the information content of an MSA. An MSA which comprises evolutionary distant relatives is considered to be highly informative that provides more discriminating conservation scores (Bartlett *et al.*, 2002a). By contrast, an MSA comprising very similar sequences is considered to be less informative. A DOPS score of 0 indicates that all positions in an alignment have the same conservation score (MSAs with low information content) while a DOPS score of 100 indicates that no two positions have the same conservation score (MSAs with high information content).

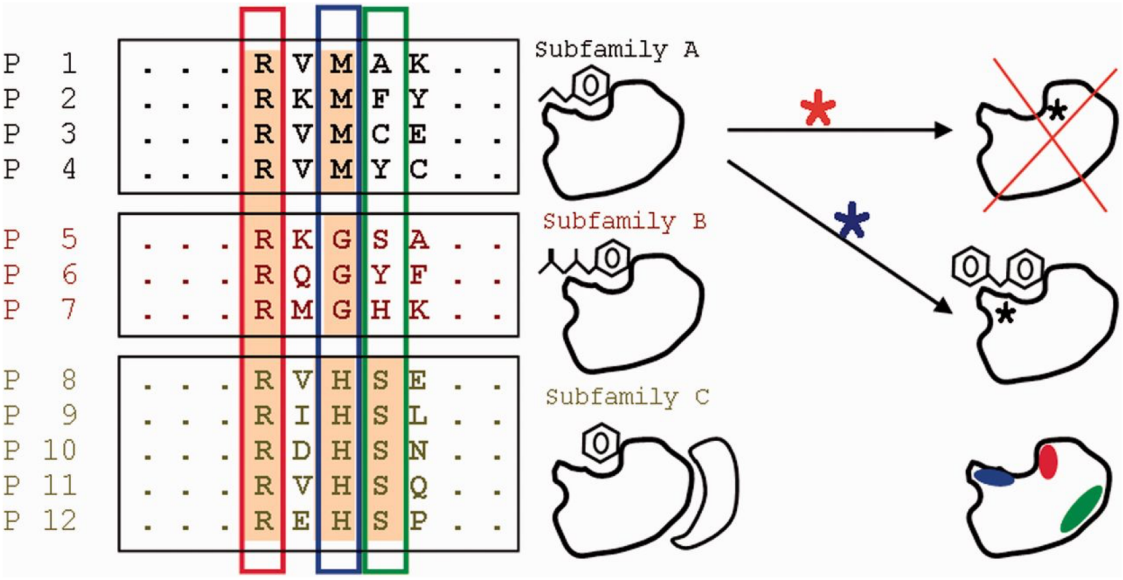


Figure 1.5: An example protein family multiple-sequence alignment (MSA) containing 12 homologous proteins that bind substrates of the same class, which can be grouped into three sub-groups or subfamilies (A, B and C) based on their substrate specificities or binding of other proteins. Structural representatives of each subfamily are also shown.The completely conserved positions and the two types of specificity-determining positions (SDPs) are highlighted in the MSA in red and blue/green respectively. These positions are mapped into a generic structural representative of the family and shown on the bottom right. For a member of subfamily A, the effects of two mutations are shown - mutation of the completely conserved position inactivates the protein while mutation of a SDP changes its substrate specificity. Taken from Chagoyen *et al.* (2015).

1.2.1.3 Identification of specificity or functional determinants

Positions that are highly conserved in a multiple-sequence alignment of homologous protein sequences are generally important for the stability, folding or common function of the proteins. By contrast, positions that are differentially conserved between different groups of homologous sequences are known as specificity-determining positions or SDPs (see Figure 1.5). These positions are known to be implicated in functional specificity of the different groups (Abhiman and Sonnhammer, 2005; Rausell *et al.*, 2010). SDPs were first studied in the Ras superfamily of small GTPases involved in various signalling functions (Casari *et al.*, 1995). In the Ras superfamily, residues involved in GTP-binding and hydrolysis are completely conserved and the functional specificity of its subfamilies generally result from different binding partners, which is reflected by the presence of SDPs involved in protein-protein interactions and in interfaces coordinating the communication between the nucleotide and membrane-binding regions (Rojas *et al.*, 2012). Casari *et al.* (1995) introduced the first method, SequenceSpace, for prediction of function determinants in proteins which represented each sequence in a protein family as a vector in a multi-dimensional space based on its amino acid sequence. This vectorial space was then reduced to a low-dimensional one, preserving most of its information using the statistical technique, Principal Component Analysis (PCA). In this low-dimensional space, vectors representing similar proteins cluster together (protein subfamilies). The method detects SDPs resides in a similar vectorial treatment for the individual residues, which generates an equivalent space where the residue clusters specific for particular subfamilies co-locate with those of the subfamilies they are SDPs for. S3Det (Rausell *et al.*, 2010) is a more recent implementation of this approach using multiple correspondence analysis (MCA). Using a systematic analysis of annotated protein subfamilies and predicted SDPs, Rausell *et al.* (2010) also reported that SDPs not only tend to accumulate in differential ligand binding sites but also in protein interaction regions indicating their possible role in the selection of interacting partners.

A number of methods have been subsequently developed for the identification of SDPs using different approaches (as reviewed in Chakraborty and Chakrabarti (2015)). The Evolutionary Trace (Lichtarge *et al.*, 1996) was another early approach used for prediction of function determinants. The ET method constructed a phylogenetic tree derived from an MSA for a query protein and its homologs. It then partitioned the phylogenetic tree into distinct branches to identify functionally similar relatives and identifies highly conserved groups of residues within homologous proteins of each branch of the tree by correlating amino-acid variations in an MSA together with structural constraints. Some methods detect SDPs in a protein family alignment, given an optimal partition of the alignment into subfamilies according with some definition of function such as enzymatic specificity by comparison of subfamily-specific sequence profiles, and analysis of relative entropy (Hannenhalli and Russell, 2000). Other approaches try to explore all possible subfamily groupings in a protein family and report that which maximizes some criterion, together with its associated SDPs. This is the strategy followed, for example, by CEO (Combinatorial Entropy Optimization) (Reva *et al.*, 2007).

GroupSim (Capra and Singh, 2008), an evolutionary rate based approach, is another widely used SDP prediction method that is used in this study. GroupSim takes an MSA containing pre-defined groups as an input and calculates a prediction score for each column in the alignment by comparing all amino acids within and between the defined groups. For each group, GroupSim first calculates the average similarity between each amino-acid pair within a group using a similarity matrix. Then it calculates, the average similarity of all amino-acid pairs between the two groups. The position-specific score of each column in the alignment is then calculated as the difference between the average within-group similarity and the average between-group similarity. The identity matrix was found to give better results than other similarity matrices. GroupSim also exploits conservation-window heuristics, which further improves its performance.

1.2.1.4 Database searching methods

With the increase in the number of annotated sequences in large databases, searching a sequence database for sequences similar to a query sequence is one of the most widely used sequence analysis tools used routinely by both bioinformaticians and biologists. The search typically provides a list of database sequences that can be aligned to the query sequence. FASTA (Pearson, 1994) and BLAST (Altschul *et al.*, 1990) algorithms have been used widely for large-scale database searches as they are much faster than dynamic programming methods described in Section 1.2.1.1.

BLAST or Basic local alignment search tool (Altschul *et al.*, 1990) begins its search for homologs by fragmenting the query sequence into short, non-overlapping sequence 'words' of length *k* (default value of *k* for proteins is 3). Using a substitution matrix, all possible words of length *k* that yield a score higher than a set threshold score when compared with one of the words in the query sequence are carried forward to compare to all the sequences in the database being searched. The database sequences are searched for matches with the words above the threshold value. The matches are subsequently extended in both directions, allowing for gaps, to generate the most significant word matches, known as high-scoring segment pairs or HSPs. This is done until the alignment score drops below a threshold or the end of either sequence is reached. The local alignments are then connected and the connected alignments are reported as a BLAST result.

The overall alignment score S for the pairwise alignment is calculated by:

$$S = (\sum M_{ij}) - gP - dG \tag{1.7}$$

where M is the score for a particular pair *ij* of residues using a substitution matrix, *g* is the number of gaps, *P* is the gap penalty, *d* is the total length of gaps and *G* is the per-residue penalty for extending the gap (Kerfeld and Scott, 2011). Since

BLAST allows the user to use different parameters such as substitution matrices, it calculates a bit score *S*' that allows comparison of alignments irrespective of different substitution matrices or gap-penalties. The bit score S' is calculated by:

$$S' = \frac{\lambda S - ln(K)}{ln(2)} \tag{1.8}$$

where *S* is the overall alignment score and λ and *K* reflect the matrices and penalties used. BLAST then calculates the Expectation value or *E*-value which reflects the probability of finding alignments with a given bit score *S*' that can be expected by chance depending on the size of a database. The E-value is calculates as:

$$E \;=\; mn\, 2^{-S'} \tag{1.9}$$

where *m* is the length of the query sequence, *n* is the total number of residues (amino acid for protein sequence searches) in the database and *S*' is the bit score. In principle, the closer the *E*-value is to 0, the better is the match. The bit scores and E-values calculated by other sequence analysis tools are calculated in a similar manner.

1.2.1.5 Sequence alignment profiles

Profile analysis is a very sensitive and specific sequence comparison method that is widely used for searching for distantly related sequences. Sequence alignment profiles capture the frequencies of each amino acid in each position of a protein MSA in a scoring table. They can be used to search a query sequence for possible matches to the profile using the scores in the table to evaluate the likelihood at each position.

The two commonly used types of sequence profiles are Position-Specific Scoring Matrices (PSSMs) (Gribskov *et al.*, 1987) and profile Hidden Markov Models (profile HMMs) (Krogh'f and Brown, 1994; Eddy, 1998). PSSMs usually contain

log-likelihood ratios instead of frequency (probability) values of each amino acid in each position of the MSA. Additionally, pairs of sequence profiles can also be compared to measure sequence similarity between two MSAs using COMPASS (Sadreyev and Grishin, 2003). COMPASS generates profiles (PSSMs) from two input MSAs followed by construction of optimal local profile-profile alignments. It then calculates an *E*-value (similar to that in BLAST) reflecting the statistical significance of detected similarities between the alignments.

HMMs are statistical models that consider all possible combinations of matches, mismatches and gaps to represent the distribution of amino acids in an alignment of a set of sequences. In a profile HMM model, for each position in an MSA, a 'match' state models the distribution of residues allowed, an 'insert' state allows for insertion of one or more residues and a 'delete' state allows the residue to be deleted. Each of these states have a table of amino acid emission probabilities, and transition probabilities for moving from state to state. The profile HMM model is built by optimizing the transition probabilities between states and the amino acid compositions of each match state in the model, such that it represents the observed amino acid variation in the alignment (Eddy, 1998). Any sequence can be represented by a path through the model such that the choice of the next state is only dependent on the choice of the current state, which is hidden. Finally, the alignment probability, given the model, is given by the product of the emission and transition probabilities along the path. HMMER (Eddy, 2009) and HHpred (Söding *et al.*, 2005) are the most widely used HMM-based packages for protein sequence analysis.

The HMMER (Eddy, 2009) software suite provides a suite of programs that can be used to create and manipulate profile HMMs and databases of profile HMMs, perform sensitive searches of sequence and profile HMM databases. HMMER is widely used for making HMM models, particularly by protein family databases such as Pfam, PANTHER, TIGRFAMS, SUPERFAMILY and Gene3D. HMMER reports both bit scores and E-values (Expectation values). The bit score

is a log-odds ratio score (base two) comparing the likelihood of the profile HMM to the likelihood of a null hypothesis (an independent, identically distributed random sequence model) and the *E*-value is calculated similar to that in BLAST.

1.2.2 Protein structure analysis

Knowledge of the three-dimensional structure of a protein plays an important role in understanding the molecular mechanisms underlying its function as it reveals the overall conformation of the protein, the biological multimeric state of the protein, binding sites, interaction surfaces and the spatial relationships of its catalytic residues. The PDBsum resource (de Beer *et al.*, 2014) provides pictorial analyses for every structure in the PDB along with detailed information extracted from various resources such as UniProtKB (UniProt-Consortium, 2015), CSA (Porter *et al.*, 2004), Pfam (Finn *et al.*, 2014), CATH (Sillitoe *et al.*, 2015) and SCOP (Murzin *et al.*, 1995), which are beneficial for protein structure-function studies. Furthermore, *ab initio* prediction of binding pockets and clefts on the protein structure can also provide useful information about protein function.

1.2.2.1 Structural alignments

Protein domains are considered to have the same fold if they share the same arrangement of secondary structure elements relative to each other in space including their relative orientation and connectivity. As protein structure is generally more conserved than sequence in evolution, distant evolutionary relationships can be captured by comparing protein folds using structure comparison methods even in the absence of any detectable sequence similarities. DALI (Holm and Sander, 1995), SSAP (Sequential Structure Alignment Program) (Taylor and Orengo, 1989) and CE (Shindyalov and Bourne, 1998) are some commonly used structural alignment methods which generate an optimal alignment of protein structures along with a score reflecting the structural similarity.

The SSAP (Sequential Structure Alignment Program) (Taylor and Orengo,

1989; Orengo and Taylor, 1996) algorithm produces a structural alignment of proteins by comparing the internal distances within each protein between proteins using double dynamic programming. SSAP originally generated only pairwise structural alignments and was later extended to produce multiple structural alignments. It has been applied in an all-against-all manner to construct the protein structure classification database, CATH. Generally, a SSAP score of at least 80 (out of 100) is associated with highly similar structures.

The structural differences between two optimally aligned structures is also commonly measured as the root mean square deviation (RMSD), calculated as Equation 1.10 between the aligned α-carbon positions.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}{\delta_i}^2} \tag{1.10}$$

In cases of high structural similarity, (i.e. a SSAP score $>$ 80 or RMSD $<$ 5 Å) functional similarity can be often suggested. However, Martin *et al.* (1998) have shown that fold similarity may not be always sufficient to conclude functional similarity as many proteins having the same function can have different folds and vice-versa.

1.2.3 Protein classification resources

Classification of proteins into homologous families has become a popular approach for functional analysis of protein sequences that provide valuable insights into our understanding of the protein function repertoire. Proteins can be classified into groups based on sequence or structural similarity. These can be then exploited for annotating uncharacterised query sequences by inheriting the annotations of proteins with known functions from the group to which the query protein is assigned.

A protein superfamily comprises evolutionarily related protein sequences that share a common ancestor. Evolutionary relationships can be either determined

by sequence, structural and/or functional similarities detected using either alignment methods or more sensitive profile searches. A family is a sub-classification of homologous proteins in a superfamily into smaller, more closely related groups according to some criteria which can vary depending on the focus of a database. For example, a sequence family at a particular level of sequence similarity groups together all proteins that share at least that level of sequence similarity, a functional family groups together homologs that share the same function and an orthologous family groups together orthologous proteins.

1.2.3.1 Sequence-based protein classifications

A large number of secondary protein databases have emerged which classify protein into families (whole protein or domain) based on locally conserved sequence patterns, known as protein signatures, which are likely to be important for structure or function. These databases use different methods for creating protein signatures such as: single motif methods (regular expressions) in PROSITE (Sigrist *et al.*, 2012), multiple-motif methods such as protein fingerprints in PRINTS (Attwood *et al.*, 2012) and full domain alignment methods in Pfam (Finn *et al.*, 2015) and SMART (Letunic *et al.*, 2015). There are a few high-quality protein family resources like PANTHER (Mi *et al.*, 2016), TIGRFAMs (Haft *et al.*, 2013) and HAMAP (Pedruzzi *et al.*, 2015) among others, which provide comparatively smaller number of manually-curated protein families. Meta-protein resources like InterPro (Mitchell *et al.*, 2014) and the Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2014) combine multiple protein family (both sequence and structure based) databases together providing higher sequence coverage compared to individual resources.

Several studies have showed that while the entire protein universe may be made up of hundreds of thousand different protein families having different multi-domain architectures, they can be generally represented by combinations of domains derived from a more limited repertoire of approximately 20,000 domain

families (Levitt, 2009; Scaiewicz and Levitt, 2015). As a result, when an uncharacterised protein does not match any characterised whole protein families, protein function can perhaps be better understood by analysing the domain components and finding homologs to each domain ('domain-based-grammar of function').

Pfam (Finn *et al.*, 2015) is a comprehensive database of 16,295 protein families that is widely used by biologists to classify and annotate proteins. In the current version (version 29.0), Pfam provides 76% coverage of the UniProtKB sequence space. For each Pfam family, a manually-curated seed alignment is created, containing a set of representative sequences that captures the diversity in the family, from which a profile HMM is generated. The profile HMM is then used to search UniProtKB to identify members of the family using strict thresholds set by the Pfam biocurators to prevent false matches. Pfam also generates higher-level groupings of related families, known as clans.

Some protein databases are built with a particular focus on specific proteins such as orthology databases that contain groups of sequences that are likely to be orthologous. The database of Clusters of Orthologous Groups (COGs) (Galperin *et al.*, 2014) provides a phylogenetic classification of proteins in complete microbial genomes where each cluster comprises of proteins that are orthologous to each other. It is widely used for function annotation of newly sequenced microbial genomes. The eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) database (Huerta-Cepas *et al.*, 2015) is a more comprehensive ortholog database than COGs and is suitable for large-scale sequence annotation of prokaryotic, eukaryotic and viral protein sequences.

1.2.3.2 Structure-based protein classifications

Several protein resources classify protein structures in a hierarchical manner in protein structure databases at the domain level on the basis of structural relationships. Because protein structure tends to be more highly conserved than sequence during evolution, structure-based superfamilies in protein structure databases

can help in bringing together more distant homologous proteins compared to sequence-based clans in Pfam. These resources take protein structures from the PDB, identify the structural domain boundaries and classify them into superfamilies based on their evolutionary origin. However, the methods used for identification of structural domain boundaries and the classification schemes differ between resources.

CATH (Orengo *et al.*, 1997) is a hierarchical protein domain classification database that is derived largely using semi-automated approaches followed by manual curation. In CATH, protein structures from the PDB are first split into separate chains. Domain boundaries in protein chains are automatically inferred if there is enough sequence or structural similarity to any domain already classified in CATH. CATHEDRAL (Redfern *et al.*, 2007), a modified version of SSAP (Taylor and Orengo, 1989) that is $\sim$1000 times faster and uses empirically derived cut-offs, is used to detect the structure similarity while the sequence similarity is determined by scanning the sequences of the new PDB chains against a library of CATH superfamily HMMs. The scores obtained from the CATHEDRAL and HMM scans are used to determine whether the domains identified in a new PDB chain are homologous to any domains classified in CATH. For domains with low similarity with domains classified in CATH, the domain boundaries are manually defined i.e. the protein chain sequence is cut at specific residues which defines the domain boundaries. The manual assignment of domain boundaries are guided by the comparison of the results of three *ab initio* domain identification methods - PUU (Holm and Sander, 1994), DETECTIVE (Swindells, 1995) and DOMAK (Siddiqui and Barton, 1995). Generally, domains that are very remote homologues have to be manually classified.

The CATH classification scheme classifies structural domains into four main levels: class (C), architecture (A), topology (T) and homologous superfamily (H) (see Figure 1.6). At the top of the hierarchy is the class level where structural domains are classified based on their secondary structure content. Within the Class

level, each domain is classified based on architecture i.e. the global arrangement of secondary structures in three-dimensional space. Each architecture is then classified into different topologies or fold groups which take into account the connectivity of the secondary structure content of the domain. Assignment of domains to topologies is done automatically using CATHEDRAL (Redfern *et al.*, 2007). Finally, within each topology or fold group, domains are classified into superfamilies according to their evolutionary origin which is based on similarities in sequence, structure and/or function. The CATH superfamily code is denoted by four numbers corresponding to each level in the CATH classification (see Table 1.2 for example).

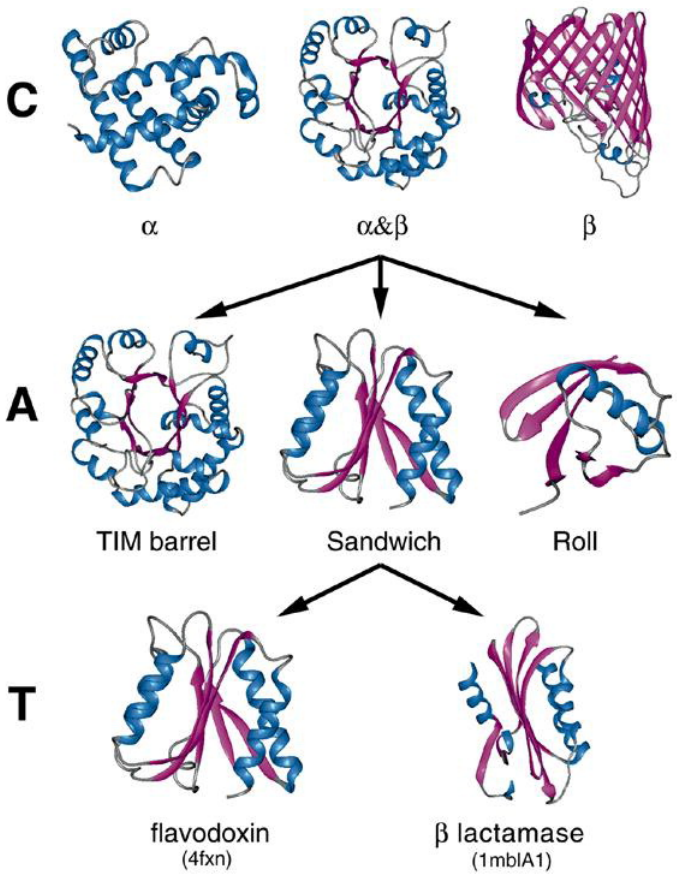


Figure 1.6: The Class, Architecture and Topology levels in CATH. Taken from Orengo *et al.* (1997).

The latest version (version 4.0) of CATH consists of 235,858 structural domains classified into 2735 homologous superfamilies (Sillitoe *et al.*, 2015). Gene3D

Table 1.2: CATH code for the DD-peptidase/beta-lactamase superfamily.

Level	CATH code	Description
Class	3	Alpha Beta
Architecture	40	3-Layer(aba) Sandwich
Topology	710	Beta-lactamase
Homologous superfamily	10	DD-peptidase/beta-lactamase superfamily

(Lees *et al.*, 2014) is a sister database to CATH which assigns protein domain sequences to the CATH superfamilies. The structural domain sequences in CATH are used to build domain superfamily-specific profile HMMs that are then used to identify domains in structurally uncharacterised protein sequences in UniProtKB (UniProt-Consortium, 2015) and Ensembl (Cunningham *et al.*, 2015).

The SCOP (Structural Classification of Proteins) (Murzin *et al.*, 1995) database classifies structural domains hierarchically by expert curation into three levels which are similar to those in CATH. The highest level of classification in SCOP is the Class based on the secondary structure content of the domain. Each class is sub-classified into Folds which bring together domains sharing the same topology. Each fold is sub-classified into Superfamilies where structures are grouped based on their common evolutionary origin. Architecture is not an explicit level in the SCOP hierarchy as in CATH but is an annotation used to describe folds. This is further sub-classified into families comprising either domains with $>$30% sequence identity or those having very similar structures and functions. The SCOP families have been found to more closely resemble taxonomic groups rather than functional groups (Pethica *et al.*, 2012). Similar to Gene3D, the SUPERFAMILY (de Lima Morais *et al.*, 2010) resource provides domain superfamily and family assignments for protein sequences based on SCOP.

SCOP2 (Andreeva *et al.*, 2014) is a successor to the SCOP database. Unlike the hierarchical classification scheme in the SCOP, the SCOP2 classification scheme uses a directed acyclic graph, where nodes form a complex network of many-to-many relationships, to organise protein structures according to structural and evolutionary relationships but additionally highlighting more complex protein

relationships.

ECOD (Evolutionary Classification of protein Domains) (Cheng *et al.*, 2014) database provides a comprehensive and regularly updated hierarchical evolutionary classification of protein domains which emphasizes more on homology compared to fold (or topology) relationships as in SCOP and CATH and focuses on remote homology. The domain classification pipeline in ECOD is automated, however, they use manual expertise for classification of proteins that do not have any confidently detectable homologs (Cheng *et al.*, 2015).

1.3 Functional diversity in domain superfamilies

Most of the highly populated domain superfamilies are universal to all kingdoms of life (Reid *et al.*, 2010). For example, the 100 most highly populated superfamilies in CATH account for $>$50% of all known domain sequences (Cuff *et al.*, 2009). These superfamilies can incorporate large amounts of structural and functional divergence during evolution even though they share a conserved structural core. Various studies (Das *et al.*, 2015a; Todd *et al.*, 2001; Reeves *et al.*, 2006; Brown and Babbitt, 2014; Galperin and Koonin, 2012) on the evolution of function in diverse protein superfamilies have illustrated the general molecular mechanisms in which a protein can gain new functions.

1.3.1 Mechanisms of functional divergence

In principle, a protein can acquire new functions by a variety of different mechanisms which may include duplication of genes, gene fusion, oligomerisation, recruitment of gene to perform a new function, post-translational modifications or alternative splicing. Various examples of the mechanisms (Figure 1.7) leading to evolution of new protein function in protein domain superfamilies are illustrated in the following section, parts of which have been published in:

Das, S., Dawson, N. L. and Orengo, C. A. (2015). Diversity in protein domain superfamilies, *Current Opinion in Genetics & Development*, 35, 40–49.

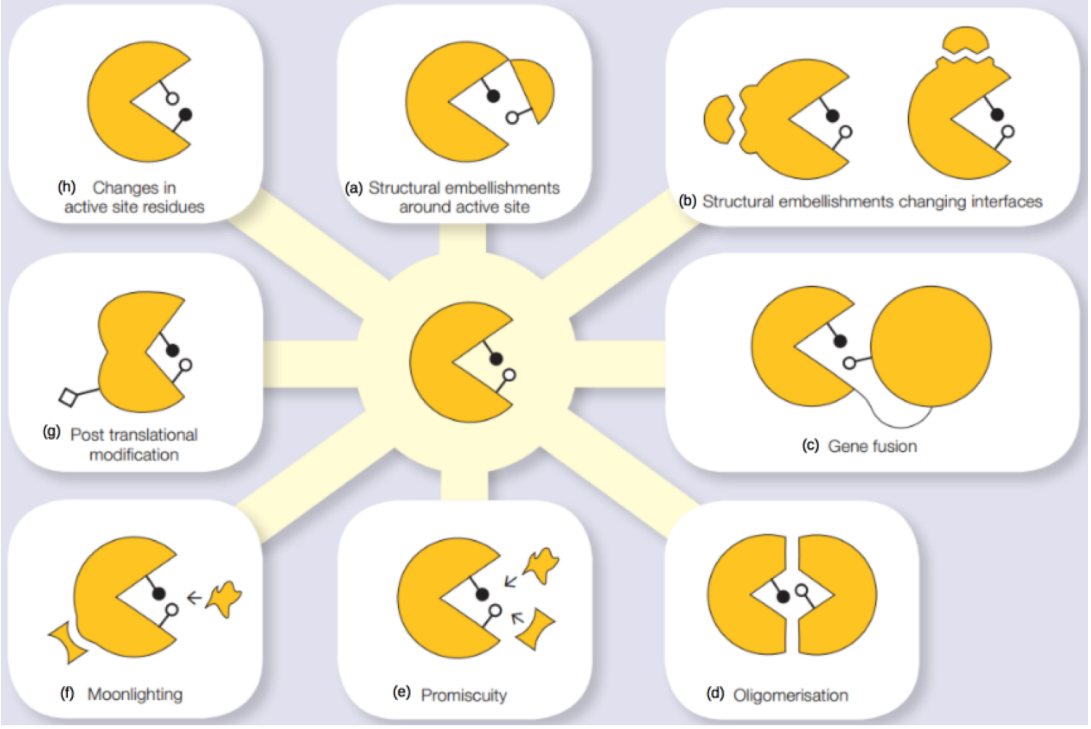


Figure 1.7: The various mechanisms, one or combination of which, can give rise to new protein functions during evolution are: (a) structural embellishments around active site, (b) structural embellishments changing interfaces, (c) gene fusion, (d) oligomerisation, (e) promiscuity, (f) moonlighting, (g) post-translational modification and (h) changes in active site residue. Note that for the mechanism panels (a), (c) and (d), one of the enzyme active site residue is contributed by its domain partner. Adapted from Das *et al.* (2015a) under CC BY 4.0.

1.3.1.1 Structural mechanisms

Homologous proteins in a superfamily share the same core domain structure, however, they can vary in size to a great extent. Structural variations within a superfamily can be due to extensive residue insertions, repeating motifs or insertions of motifs and these structural changes are often accompanied by modifications in function (Cuff *et al.*, 2009; Sandhya *et al.*, 2009). For example, for more than 150 CATH superfamilies accounting for half of all known domains, at least a two-fold variation in the size is observed between the most diverse domains (Reeves *et al.*, 2006), however, the structural core of the domain is highly conserved even for distant relatives (see Figure 1.8).

Very extensive residue insertions generally adopt secondary structures fea-

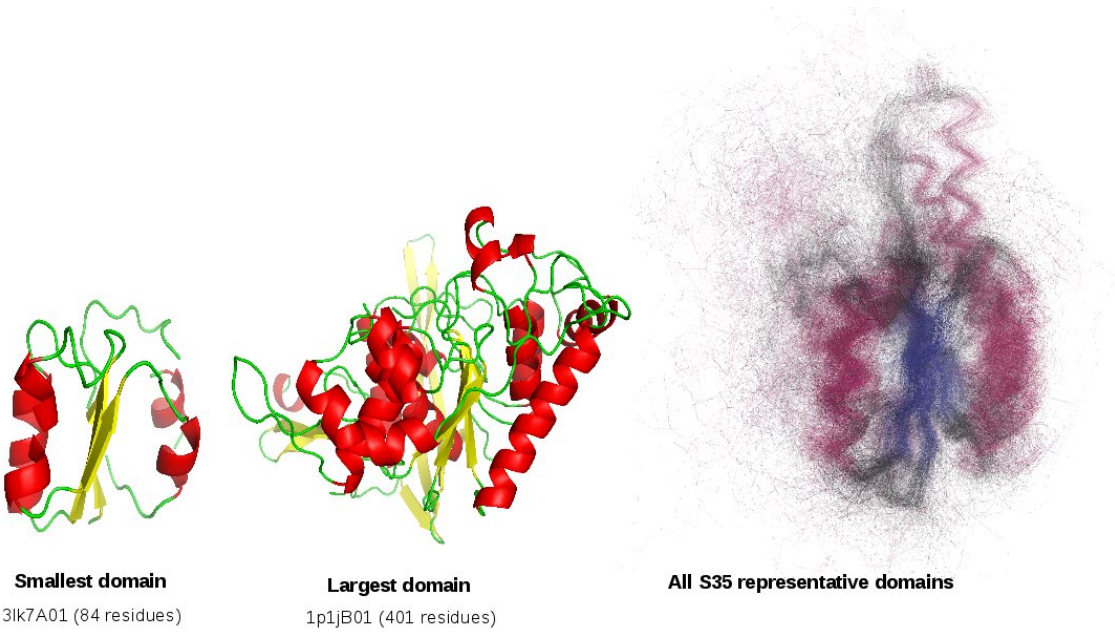


Figure 1.8: Structural diversity in the NAD(P)-binding Rossmann-like superfamily in CATH (CATH 3.40.50.720). The structures of the smallest and largest domain in the superfamily along with the superposition of all non-redundant domains at 35% sequence identity (S35 representative domains) is shown to highlight the conserved structural core. Taken from Das *et al.* (2015a) under CC BY 4.0.

tures that form structural decorations to the domain core and are frequently located close to functional sites where they can modify active site geometry by diversifying substrate specificity or altering surface features changing protein interaction partners. For example, in the HUP (High-signature proteins, UspA, and PP-ATPase) superfamily, the structural diversity among the domains can range from small changes close to the active site to extensive embellishments which mediate changes in molecular function affecting the biological processes in which they are involved (Dessailly *et al.*, 2010). Similarly, a recent large-scale activity profiling of the haloalkanoic acid dehalogenase (HAD) superfamily revealed a high degree of substrate ambiguity among the superfamily members and suggested that domain insertions to the core catalytic Rossmann fold may drive the evolution of new functions i.e increased substrate range in this superfamily (Huang *et al.*, 2015).

1.3.1.2 Molecular tinkering

The catalytic machinery may be highly conserved in many highly diverse superfamilies, for example, diverse enzymes such as peptidases, thioesterases, lipases of the α/β hydrolase superfamily utilize the same catalytic triad Ser-His-Asp for different types of bond changes (Todd *et al.*, 2001). However, changes in the catalytic apparatus among enzymes in other superfamilies can provide a huge diversity of functions as even small changes associated with residue mutations in a binding or active site can result in alteration of the shape, physico-chemical and electrostatic characteristics of the site significantly. A recent study has indicated a significant number of highly diverse enzyme superfamilies in CATH which show a wide range of diversity in both changes in catalytic residues and changes in position of catalytic residues in the protein scaffold by different members of the superfamilies (Furnham *et al.*, 2015). For example, members of the enolase superfamily, having a TIM-barrel fold, show a huge diversity of enzymatic activities by using different sets of catalytic residues (Figure 1.9) (Wichelecki *et al.*, 2014). Similarly, relatives in the diverse protein kinase-like (PKL) superfamily share ten

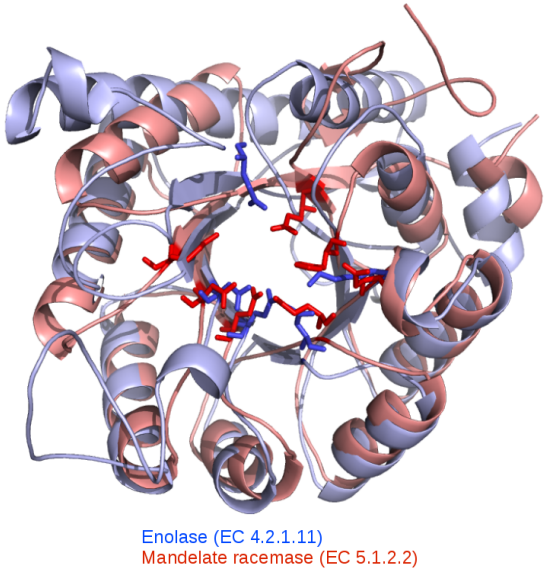


Figure 1.9: Functional diversity by molecular tinkering of the catalytic domain in Enolase (shown in blue) and Mandelate racemase (shown in red) of the Enolase superfamily. The active site residues of each of the protein domains are shown as sticks.

key residues conserved across the entire superfamily. However, there is considerable diversity of the active sites among different families, with nine out of ten key residues substituted in at least one family in the superfamily (Kannan *et al.*, 2007). On the contrary, significantly different catalytic machineries in some enzyme superfamilies have highly similar functions and substrates, suggesting either convergence within the superfamily or evolutionary drift from a common functional ancestor along different routes.

In addition to changes in the active or binding site, mutations of residues in protein-protein interfaces can also occur conferring diverse specificities in different functional relatives within a superfamily (Dessailly *et al.*, 2013). A large scale study of 645 functionally diverse CATH superfamilies reported that the cumulative binding sites from diverse relatives covered most of the protein domain surface and were associated with a wide range of protein partners (Dessailly *et al.*, 2013). In contrast, sometimes the same protein surface is exploited, but by different partners. For example, the diversity in the reactions carried out by the enzymes of the two dinucleotide-binding domains flavoproteins (tDBDF) superfamily is achieved by different protein partners acting as electron acceptors and interacting with the same protein surface of the tDBDF domains (Ojha *et al.*, 2007; Dessailly *et al.*, 2013).

1.3.1.3 Different multi-domain contexts

Changes in the multi-domain architecture (MDA) of a protein can significantly alter its context, thereby modifying its function. For example, the TIM barrel glycosyl hydrolase domain can exist in different domain organizations that modulates its substrate specificity (Todd *et al.*, 2001). Also, in the highly diverse Thiamine pyrophosphate(TPP)-dependent enzymes, which catalyse a large number of different reactions using TPP as the co-factor, changes in domain partnership alters the size and physico-chemical properties of the active site pocket, leading to a huge range of substrates, products and stereo-selectivity (Vogel and Pleiss,

2014).

TPP-dependent enzymes bind TPP at the interface of two conserved homologous domains: the pyrophosphate (PP) domain and the pyrimidine (PYR) domain. The PP and PYR domains have very low sequence identity but due to their very high structural similarities they are thought to have diverged from a common ancestor through gene duplication and then fused into a single chain (see Figure 1.10).

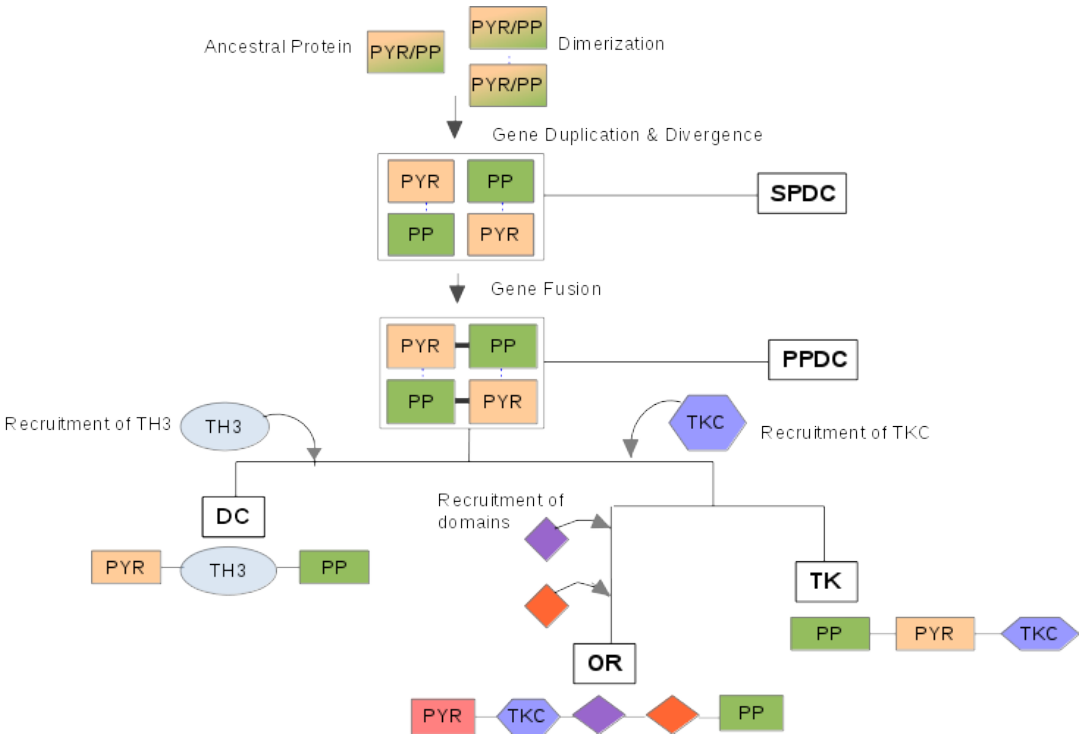


Figure 1.10: Evolutionary history of the Thiamine pyrophosphate (TPP)-dependant enzyme superfamily (CATH 3.40.50.970). The superfamily comprises two conserved homologous domains: the pyrophosphate (PP) domain and the pyrimidine (PYR) domain. TPP-dependent enzymes have been classified broadly into: DC (decarboxylases), TK (transketolases), OR (oxidoreductases), SPDC (sulfopyruvate decarboxylase) and PPDC (phosphopyruvate decarboxylase) (Costelloe *et al.*, 2008).

Different oligomerisation states also effectively change the domain context. Again, multiple oligomerisation states have evolved in different species in the TPP-dependent superfamily. Whilst some may be associated with enhanced stability, others clearly influence active site characteristics by changing the positioning of domains providing catalytic residues (see Figure 1.11).

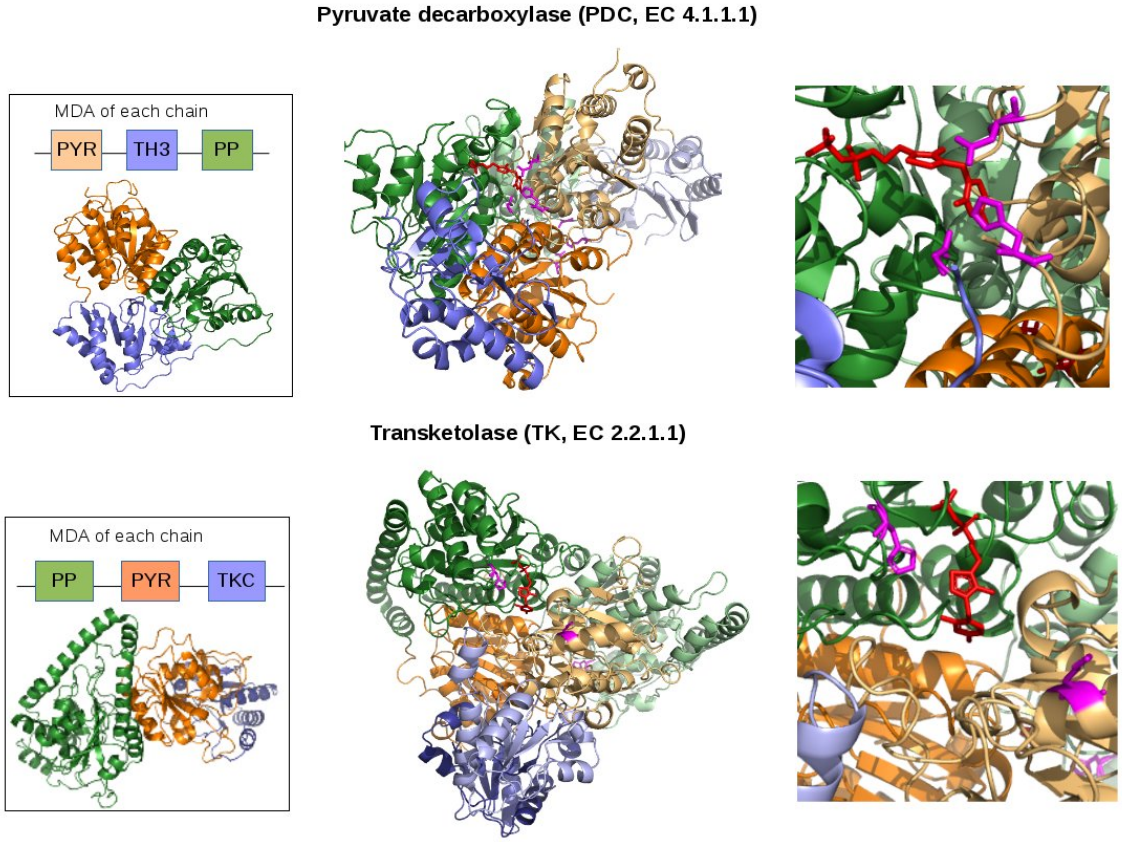


Figure 1.11: Functional diversity in the Thiamine pyrophosphate (TPP)-dependent enzyme superfamily due to changes in domain contexts. Pyruvate decarboxylase (PDC, EC 4.1.1.1) and transketolase (TK, EC 2.2.1.1) in the TPP-dependent superfamily both consist of two chains comprising two TPP domains PP and PYR (chains are represented by darker and lighter shades of each constituent domain colour). The left hand images show the difference in multi-domain architectures and 3D arrangements for these two proteins. The middle images show the different dimeric assemblies that the proteins form. The right images zoom in on the active sites. The TPP molecule is shown in red and the catalytic residues are shown in magenta. Catalytic residues are contributed from the PP domain of one subunit and the PYR of the other subunit. In TK the size of the active site pocket is larger. Taken from Das *et al.* (2015a) under CC BY 4.0.

1.3.1.4 Promiscuity and moonlighting

Functional diversity can also be observed between closer homologs, and sometimes within the same protein in different contexts. For example, proteins can have multiple catalytic activities as in promiscuous and moonlighting proteins. Promiscuous enzymes are widely accepted as starting points of evolution of new functions in a superfamily (Pandya *et al.*, 2014). Often, a promiscuous ancestral protein, under natural selection, can give rise to a specialist enzyme using a variety of different mechanisms like rearrangements in the catalytic metal ions, binding of alternative cofactors and domain insertions. For example, recent studies on the Metallo-beta-Lactamase superfamily showed that the high degree of promiscuity among its enzymes is due to the plasticity of the catalytic metal ion-binding sites (Baier and Tokuriki, 2014) and use of alternative metals as cofactors for the metalloenzymes (Baier *et al.*, 2015). On the other hand, moonlighting proteins perform completely different functions to their native activity (Jeffery, 1999). However, they do not have a common mechanism through which they switch between different functions and orthologous proteins in different organisms do not necessarily share the moonlighting functions (see Chapter 5 for details).

1.3.1.5 Combination of mechanisms

In majority of the large functionally diverse superfamilies, functional diversity generally results from a combination of different molecular mechanisms, some of which have been discussed above. For example, functional diversification in the PD-(D/E)XK phosphodiesterase superfamily is attributed to structural embellishments to the core, domain swapping, active site residue variation and changes in multi-domain architecture (Steczkiewicz *et al.*, 2012).

Apart from the mechanisms mentioned here, there are diverse post-transcriptional and post-translational mechanisms that further expand the functional diversity of proteins which do not necessarily affect the canonical sequence or structure of the protein. Major post-transcriptional modifications include alternative splicing

and RNA editing while post-translational modifications include phosphorylation, glycosylation, methylation and acetylation among others. Many proteins also contain short linear motifs (SLiMs) comprising generally of 3-11 contiguous amino acids in intrinsically disordered regions that are involved in signalling, regulatory function and mediating protein-protein interactions, often independent from other functions of the proteins in which they occur.

1.3.2 Capturing diversity in superfamilies

By classifying proteins into superfamilies and families, and bringing together information on their sequences, structures and functions, one can explore how function is modulated by sequence and structural changes. Such insights are not only essential for inheriting annotations between homologous proteins in order to cope with the huge dearth in experimental annotations but also for understanding the effect of genetic variations on protein function. If all protein sequences in the public databases like UniProtKB are functionally classified into families, it would be possible to at least detect the subtle changes in conservation patterns that can suggest shifts in binding specificities or catalytic machineries. These data can also guide experiments to focus on unusual relatives to more comprehensively landscape the functional repertoires of all protein superfamilies.

1.4 Overview of thesis

Chapter 2 describes the development of a new algorithm, FunFHMMer, for functional classification of CATH protein domain superfamilies and identification of functional families (or FunFams) using evolutionary signals in sequence alignments.

Chapter 3 discusses the automated function prediction pipeline and web server which exploits the CATH FunFams generated by FunFHMMer and its benchmarking using in-house datasets and the Critical Assessment of Protein Function Annotation (CAFA) 2 benchmark.

Chapter 4 provides an in-depth analysis of the FunFams generated by FunFHMMer and investigates how the quality of the FunFams can be improved. The utility of the FunFams is then examined for exploring superfamily diversity in CATH and identifying functional sites in proteins.

Chapter 5 proposes a classification scheme based on the structure-function analysis of selected moonlighting proteins and discusses the use of the FunFHMMer function prediction pipeline for annotation of moonlighting proteins.

Chapter 6 closes this thesis with an overall summary of the work and future directions for the FunFHMMer pipeline. Several strategies are proposed to improved the quality of the FunFams and the FunFHMMer pipeline and final remarks are presented on the possibilities the domain-based approach developed in this work provides in future for making sense of the huge amount of biological data available.

Chapter 2

FunFHMMer: functional classification of domain superfamilies

2.1 Background

Homologous proteins can often evolve different functions (see Section 1.3 in Chapter 1) as a result of different sets of residues in their active site, addition of secondary structure embellishments to the core protein structure which alters the geometry of the active site of the protein or an interface on the protein or due to domain-shuffling in multi-domain proteins which can alter the context of the domain and again result in changes to functional sites. One way to capture and understand this functional diversity among protein homologs is to functionally classify protein superfamilies.

Classification or clustering of the known parts of the protein universe into functional groups can not only help in predicting functions of proteins but can also provide valuable insights into how the protein function repertoire evolves. As well as increasing the accuracy of functional inheritance between relatives, such functional grouping would also facilitate multiple sequence alignment of the relatives to find conserved residue positions which can provide valuable insights about the key functional sites and mechanisms of the protein.

2.1.1 Clustering methods for protein sequences

All protein clustering methods require an all-against-all sequence similarity matrix of a dataset generated by sequence analysis tools such as BLAST (Altschul *et al.*, 1990). Clustering of protein sequences to give a meaningful functional classification is a non-trivial task because of two main reasons: firstly, sequence similarity does not always perfectly correspond to functional similarity and secondly, due to domain-shuffling. The clustering algorithms that are commonly applied to protein

sequence datasets can be broadly grouped into hierarchical, partitioning, graph-based and greedy incremental approaches.

2.1.1.1 Hierarchical clustering

Hierarchical clustering methods group data objects into a tree of clusters (dendogram) by performing all-against-all comparisons between data objects either in a bottom-up fashion i.e. from the leaf nodes to the root (agglomerative hierarchical clustering) or in a top-down fashion i.e. from the root to the leaf nodes (divisive hierarchical clustering). A clustering of the data objects is later obtained by cutting the dendogram at a desired similarity or granularity level.

A large number of protein sequence clustering algorithms make use of agglomerative hierarchical clustering such as CluSTr (Petryszak *et al.*, 2005), SYSTERS (Krause *et al.*, 2005), ProtoMap (Yona *et al.*, 2000) and ProtoNet (Rappoport *et al.*, 2013). All of these methods use a BLAST-based sequence similarity measure for the clustering, however, the implementation, similarity measures and the level of granularity is different for each algorithm (Liu and Rost, 2003). ProtoNet (Rappoport *et al.*, 2013), the only algorithm out of these which is actively maintained, provides an unsupervised agglomerative hierarchical clustering of all proteins in the UniProtKB database at all levels of cluster granularity (i.e. from singletons to root clusters whose proteins share no apparent similarities) and associates annotations to the proteins from all leading annotation resources for structure, sequence, function and taxonomy.

2.1.1.2 Partitioning clustering

In partitioning clustering, data objects are initially partitioned into a fixed number of clusters and during the clustering process, data objects relocate from one cluster to another based on their similarity to the closest cluster. K-means, k-medoids, k-modes and PAM (Partitioning Around Medoids) are widely used partitional clustering algorithms (Fayech *et al.*, 2009).

Partitioning methods that have been used to cluster protein sequences include a scalable algorithm (Guralnik and Karypis, 2001) for clustering sequential data based on a k-means approach, JACOP (Sperisen and Pagni, 2005) which is based on a random sampling of sequences into groups exploiting the PAM partitioning algorithm and an information-theoretic entropy-driven partitioning algorithm (Rappoport *et al.*, 2014) for the hierarchical ProtoNet (Rappoport *et al.*, 2013) tree (see Section 2.1.1.1). The information-theoretic partitioning approach (Rappoport *et al.*, 2014), which finds an optimal partitioning of the ProtoNet tree by minimizing the entropy-derived distance between annotation-based partitions and all available hierarchical partitions, was found to be superior to a naive cut of the tree.

2.1.1.3 Graph-based clustering

Graph-based clustering methods generate fully connected graphs from an all-against-all comparison of data objects in which the nodes represent the data objects and the edges represent the distance between them. Edges representing distances below a certain fixed threshold are removed and a clustering obtained by grouping data objects that are still linked. When clustering protein sequences, the graph to be clustered is a sequence similarity network, in which the nodes represent sequences and the edges represent sequence similarity relationships between them.

The graph-based algorithms that are most widely used to cluster protein sequences are Markov clustering (MCL) (Van Dongen, 2000), affinity propagation clustering (Frey and Dueck, 2007) and spectral clustering (Paccanaro *et al.*, 2006). MCL (Van Dongen, 2000) is based on the principle that random walks on a graph rarely connects one natural cluster (presence of many edges between the members of that cluster) to another. MCL simulates flow within a graph by creating a stochastic Markov matrix of transition probabilities between all sequences from an all-against-all sequence similarity matrix and promotes flow in a highly connected

region (matrix expansion) and restraining the flow (matrix inflation) in sparsely connected regions, thus revealing the natural clusters within the graph. TRIBE-MCL (Enright *et al.*, 2002) is an extension of MCL for generating protein families based on precomputed sequence similarities using BLAST and the inflation value parameter of the MCL algorithm is used to control the granularity of the families.

2.1.1.4 Greedy incremental clustering

Greedy incremental clustering methods are frequently used to cluster large protein sequence datasets to scale up the clustering in a fast but heuristic manner (Hobohm *et al.*, 1992). CD-HIT (Fu *et al.*, 2012), kClust (Hauser *et al.*, 2013) and UCLUST are some of the widely used heuristic protein clustering programs. These methods first sort all sequences by length. The longest sequence is taken as the representative of the first cluster. Then the remaining sequences are taken as query and compared with the representative sequences of the already created clusters. If the query sequence fulfils the method's similarity criteria with one of the representative sequences, the query sequence is added to that cluster, otherwise a new cluster is created for which the query becomes the representative sequence. The methods use BLAST-like short word filtering to determine the similarity between two sequences rather than performing an actual sequence alignment. CD-HIT (Fu *et al.*, 2012) is routinely used in sequence analysis to reduce sequence redundancy and for various other applications. However, while CD-HIT can be used to cluster large sequence datasets down to 50% sequence identity, kClust (Hauser *et al.*, 2013) has been developed to cluster at 20%-30% maximum pairwise sequence identity.

2.1.2 Automated classification of protein families

Several clustering methods (see Section 2.1.1) have been used for automated classification of widely-used protein resources, some of which are described briefly below.

ADDA (Automated Domain Delineation Algorithm) (Heger and Holm, 2003) is a clustering algorithm that identifies protein domain families within large protein sequence datasets and was used to build Pfam-B families (Finn *et al.*, 2014). Firstly, ADDA computes all-vs-all pairwise alignments with BLAST. It identifies domain boundaries within protein sequences by locating where BLAST alignments are located on a sequence and splits the sequence such that a minimum number of alignments are cut by domain boundaries and that a maximum number of alignments stretch over complete domains followed by an iterative refinement of the domain boundaries using an optimization strategy. Finally, the domain sequences are arranged in a minimum spanning tree where the similarity between two domain sequences is determined by their relative overlap given a BLAST alignment. Spurious links in the tree are then removed by checking pairwise profile-profile comparisons of adjacent domain sequences. The remaining connected domain sequences are then taken to represent protein domain families.

PhyloFacts (Krishnamurthy *et al.*, 2006), a phylogenomic encyclopedia of protein families across the Tree of Life, classifies its families into subfamilies using the SCI-PHY algorithm (Sjolander, 1998). The SCI-PHY (Subfamily Classification in Phylogenomics) algorithm combines agglomerative hierarchical clustering with an unsupervised clustering evaluation strategy to identify protein families in an *ab-initio* manner. During the clustering process, SCI-PHY generates residue distribution profiles to represent the clusters and uses the relative entropy between these profiles as the cluster dissimilarity measure. During this process, it uses a residue probability density function in the form of a Dirichlet mixture density (Sjölander *et al.*, 1996) derived from the residue distributions observed in sets of high-quality alignments in the BLOCKS database (Henikoff and Henikoff, 1992). The use of Dirichlet mixture densities helps to create more specific and selective profiles than those based on common substitution matrices (Brown *et al.*, 1993). Secator (Wicker *et al.*, 2001) is another phylogenomic subfamily identification method which uses a sequence dissimilarity measure in order to cut

a phylogenetic tree. These methods invariably require an accurate multiple sequence alignment of the protein family as a starting point in their pipeline which is likely to be erroneous for very large and very diverse families.

2.1.3 Quality assessment of automated classification methods

Quality assessment of automated protein classification methods is essential for both computational and experimental biologists to know whether the automatically generated protein families correspond well with manually-curated, experimentally characterised protein families and to decide whether the automated classification provided by a method can be relied upon. Pfam (Finn *et al.*, 2014) families are often used as reference families for benchmarking automated classification methods, however, it is a non-trivial task to choose the families for benchmarking.

Specialized manually-curated resources like the Structure-Function Linkage Database (SFLD) (Brown *et al.*, 2006) focuses on hierarchical classification and provides insights into the structure-function relationships of few functionally or mechanistically diverse enzyme superfamilies. SFLD has been previously used as a benchmark dataset for validation of superfamily classification methods such as SCI-PHY (Brown *et al.*, 2007) (see Section 2.1.2) and GeMMA (Lee *et al.*, 2010). Other specialised curated resources include the Thiamine pyrophosphate (TPP)-dependent Enzyme Engineering Database (TEED) (Widmann *et al.*, 2010) and the Carbohydrate-Active Enzymes (CAZy) database. The TEED database provides a manually curated classification of 9 TPP-dependent enzyme superfamilies which are sub-classified into families based on sequence similarities (Vogel and Pleiss, 2014). TEED was established by a BLAST search against the NCBI sequence database for 62 experimentally verified TPP-dependent enzymes (seed sequences) followed by assignment to protein families on the basis of their sequence similarities followed by manual curation based on domain organization and other criteria. Similarly, the CAZy database provides a comprehensive cu-

rated sequence-based family classification of enzymes that are involved in the synthesis, degradation and modification of carbohydrates. The CAZy families are seeded using experimentally characterized proteins, and are then populated by sequences from public databases with significant similarity (Henrissat, 1991; Cantarel *et al.*, 2009). The functional and structural information of the database is curated on a regular basis.

2.1.3.1 Structure-Function Linkage Database (SFLD)

The Structure-Function Linkage Database (SFLD) (Brown *et al.*, 2006) provides a manually-curated, gold standard set of mechanistically diverse enzyme superfamilies classified into families based on the reactions catalysed by the enzymes. SFLD currently contains 12 gold-standard core superfamilies, only 9 have been classified into families, namely Amidohydrolase, Crotonase, Enolase, Haloacid dehalogenase, Isoprenoid Synthase Type I, Isoprenoid Synthase Type II, Nucleophilic Attack 6-Bladed Beta-Propeller (N6P), Radical SAM and Rubisco.

Evaluation measures

The performance of a family-identification protocol can be assessed on the SFLD superfamilies benchmark by using a single performance score (see Equation 2.1) used by Lee *et al.* (2010) which incorporates three distinct measures: purity, edit distance and VI distance introduced by Brown *et al.* (2007), which captures the desired balance between high sensitivity and high specificity.

The following equations and large parts of the explanatory text are thus directly taken from these publications:

Purity. Purity is defined as the percentage of families within which all annotated members are annotated with the same function. It is calculated as:

$$Purity(p) \ = \ \frac{Number\ of\ pure\ families}{Total\ number\ of\ families} \ . \ 100 \tag{2.1}$$

Edit distance. Edit distance can be defined as the number of split or merge operations that are required to transform the predicted families into the families that correspond to the experimental functional annotations. The edit distance between the true partition (S) and the predicted partition (S') with families k and k' respectively, is calculated as:

$$Edit\ distance(e) \ = 2\{\sum_{k,k'} r_{k,k'}\} - K - K' \tag{2.2}$$

where $r_{k,k'} \ = \ 1$ if families have sequences in common otherwise $r_{k,k'} \ = \ 0$ and K and K' are the total number of families in S and S'.

VI distance. VI distance can be defined as the amount of information not shared between the predicted families and the experimentally annotated families. The VI distance between S and S' is calculated as,

$$VI\ distance(v) = H(S) + H(S') - 2I(S, S') \tag{2.3}$$

where H is the entropy of a partition and I is the mutual information between two partitions,

$$H(S) = \sum_{k=1}^{K} \frac{n_k}{N} log \frac{n_k}{N} \tag{2.4}$$

$$I(S, S') = \sum_{k=1}^{K} \sum_{k'=1}^{K} \frac{n_{k,k'}}{N} log \frac{n_{k,k'}}{N} \tag{2.5}$$

In Equations 2.4 and 2.5, n_k is the number of items in the family k of partition S, $n_{k,k'}$ is the number of overlapping items between the family k in partition S and the family k' in partition S' , and N is the total number of items in the set. Identical partitions will have both an edit and VI distance of zero.

Performance Score. The performance of the family-identification protocols on the SFLD superfamilies benchmark were compared using a performance score (Lee *et al.*, 2010) incorporating three distinct measures described above. The performance score (range 0-100) is then calculated as,

$$Performance\ score = \frac{2p\ +\ (100 - c_e.e)\ +\ (100 - c_v.v)}{4} \tag{2.6}$$

where e_0 and v_0 are the initial values of edit and VI distance respectively and $c_e = \frac{100}{e_0}$, $c_v = \frac{100}{v_0}$.

2.1.4 Functional classification of CATH superfamilies

The expansion of CATH superfamilies with increasing amount of sequence data (Lees *et al.*, 2014) and functional data (Bairoch, 2000; Ashburner *et al.*, 2000), have provided a wealth of data for large-scale studies of functional divergence within superfamilies (Todd *et al.*, 2001; Cuff *et al.*, 2009; Das *et al.*, 2015a). These studies have shown that for the majority ($>$ 90%) of the superfamilies in CATH-Gene3D, the sequence relatives have highly similar structures and functions. However, these conserved superfamilies tend to be small and highly specific to certain species or sub-kingdoms of life. In the remaining universal and highly populated protein superfamilies ($<$ 5% of CATH superfamilies accounting for $>$ 50% of all domains), there is significant divergence of function between relatives because of which, one of the major challenges of using the CATH superfamilies for functional annotation is the sub-classification of relatives in these superfamilies into coherent functional groups. This prompted the development of automated protocols to sub-classify the CATH superfamilies into functional families i.e. families comprising domain sequences sharing the same function or sub-function within a superfamily.

2.1.4.1 GeMMA algorithm for clustering domain sequences

GeMMA (Genome Modelling and Model Annotation) (Lee *et al.*, 2010) is an agglomerative hierarchical clustering algorithm which clusters protein domain sequences in CATH-Gene3D superfamilies.

For each CATH superfamily, the protein domain sequences from the corresponding Gene3D superfamily are first pre-clustered at 90% sequence identity into S90 clusters using CD-HIT (Fu *et al.*, 2012). Fragments (sequences having a length less than 80% of the average sequence length of the cluster) are then removed from the remaining clusters, which form the starting clusters for GeMMA to build a bottom-up clustering tree (GeMMA tree) from the leaf nodes to the root (see Figure 2.1). Multiple sequence alignment are built for each cluster using MAFFT (Katoh *et al.*, 2002). GeMMA then exploits the COMPASS (Sadreyev and Grishin, 2003) algorithm to compare the sequence profiles derived from multiple sequence alignments of pairs of clusters present at each iteration of the clustering. After each iteration, the cluster profiles matching above a certain threshold are merged and alignments are generated for the new clusters. These iterations continue till a single cluster remains, generating the bottom-up hierarchical clustering tree. The hierarchical clustering in GeMMA is similar to that used in the multiple sequence alignment method, MultAlin (Corpet, 1988) to determine the order in which sequences should be compared and gradually aligned to generate an optimal multiple sequence alignment using the scores of all pairwise sequence comparisons as an index of similarity between the sequences.

GeMMA uses the E-value corresponding to the profile-profile comparison of pairs of clusters (i.e. child nodes in the emerging tree) using COMPASS to monitor the progress (iteration) of the algorithm. As the E-values increase from the leaf nodes (i.e. the S90 starting clusters) towards the root, the sequence clusters become more diverse in their sequences and functions. Since GeMMA is a computationally expensive algorithm, two heuristics - greedy merging and comparison sampling - were implemented to speed up the clustering of highly populated su-

perfamilies in the DFX algorithm (Rentzsch, 2012, described in the next section). This was reported to provide the same performance level as the original GeMMA implementation.

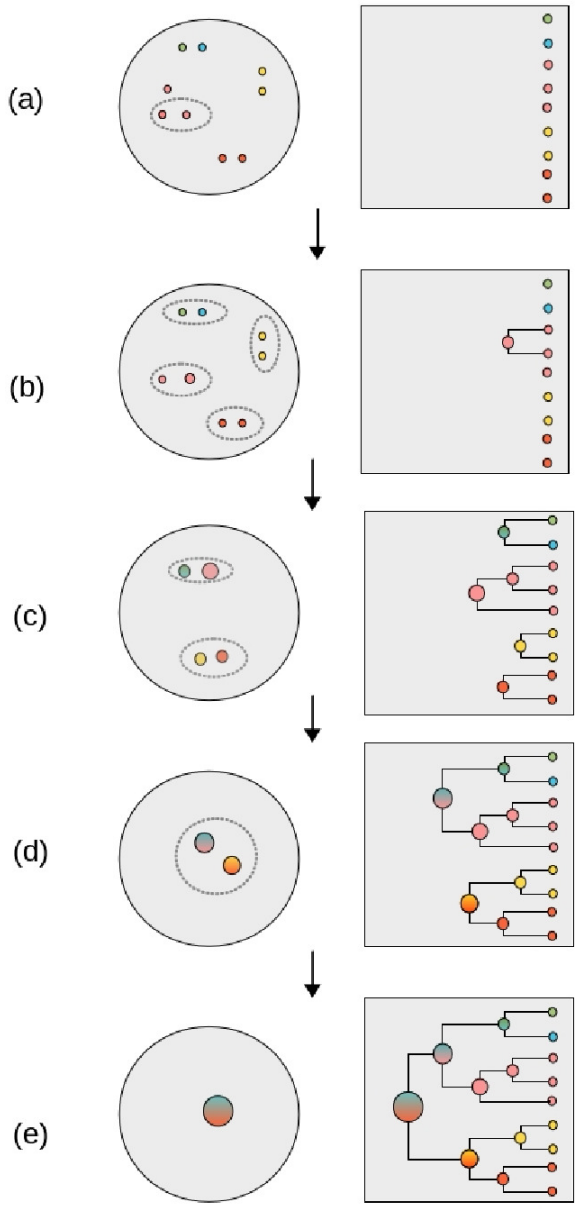


Figure 2.1: Hierarchical agglomerative clustering of sequences in a CATH-Gene3D superfamily by GeMMA (Lee *et al.*, 2010). The coloured circles represent the sequence clusters where each colour denotes a unique function. **(a)** The S90 clusters form the starting clusters for GeMMA. **(b)-(e)** The cluster merges (shown inside the grey circles) are traced by the GeMMA clustering tree (shown inside the grey boxes) till a single cluster is formed.

The aim of clustering the superfamily domain sequence data using GeMMA is to partition the resulting clustering tree ideally for all superfamilies into sepa-

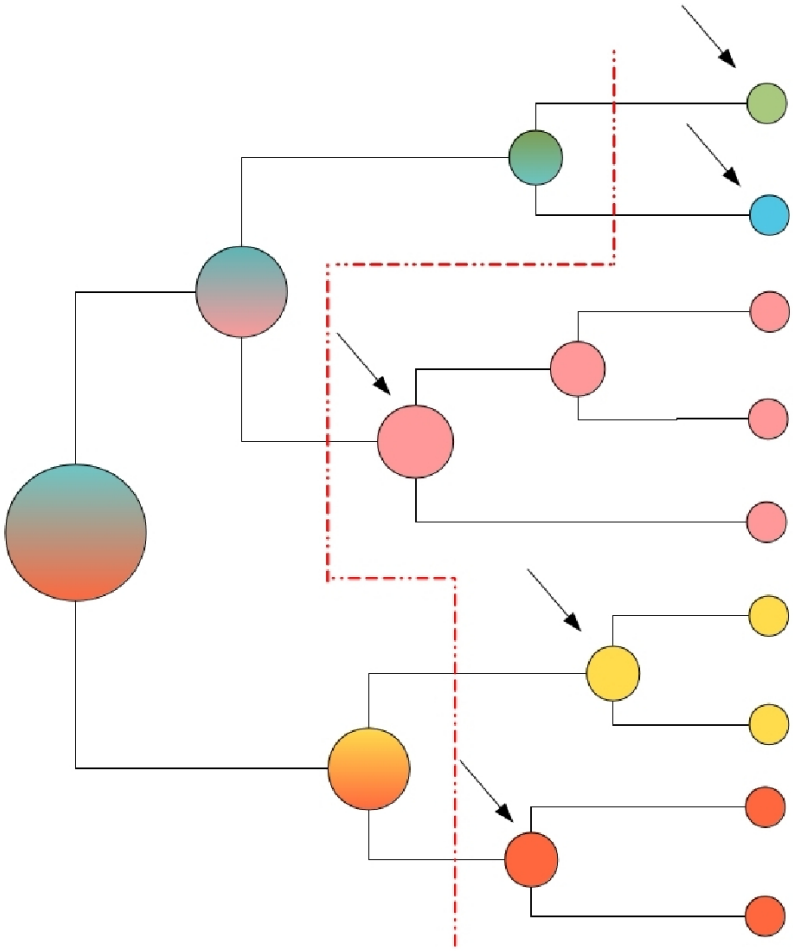


Figure 2.2: Ideal partitioning of GeMMA tree. The coloured circles represent the sequence clusters where each colour denotes a unique function and the red dashed line denotes the optimal cut of the GeMMA tree. The ideal functional families are indicated by arrows. Taken from Das and Orengo (2016) under CC BY 4.0.

rate clusters of sequences performing different functions such that the sequence relatives in each cluster share the same function (see Figure 2.2). However, finding an optimal partitioning of a hierarchical tree of sequence relatives is not trivial. GeMMA classifies the sequences in each CATH superfamily into families by cutting the GeMMA tree at a generic granularity threshold of $E = 10^{-40}$ (Lee *et al.*, 2010). The performance of GeMMA was found to be comparable to that of SCI-PHY (Lee *et al.*, 2010).

2.1.4.2 DFX protocol for functional classification

Family identification in CATH superfamilies was later improved by a supervised family identification algorithm, DFX (Domain Family Exploration algorithm) (Rentzsch and Orengo, 2013). The starting point of the DFX pipeline is a hierarchical tree of sequence relatives, i.e. the GeMMA clustering tree for each CATH superfamily. Firstly, DFX associates each cluster node in the tree with a set of high-quality GO annotations from UniProt-GOA (Dimmer *et al.*, 2012), that are associated with the annotated sequences of the cluster. GO annotations are considered to be of high-quality if they have: Inferred from Electronic Annotations (IEA) in Swiss-Prot made by either EC2GO or Swiss-Prot Keyword2GO mapping methods, as well as experimentally inferred or curated (non-IEA) annotations in UniProtKB (Rentzsch and Orengo, 2013). The IEA GO annotations in Swiss-Prot from EC2GO or Swiss-Prot Keyword2GO mapping methods were included since they primarily represent a description of the Enzyme Commission (EC) annotations or manually-derived Swiss-Prot Keywords in terms of GO annotations (Škunca *et al.*, 2012). Secondly, DFX removes any cluster which lacks at least one sequence with high-quality GO annotations. Finally, DFX assesses the functional coherence of all cluster nodes in the GeMMA tree based on GO annotation data associated with the sequences in each cluster node and removes all nodes that are not judged as functionally coherent. This results in partitioning of the tree into functional families (FunFams) for each CATH superfamily.

An analysis of 466 enzyme superfamilies having full Enzyme Commission number (EC4) annotations showed that the functional families (FunFams) generated by DFX were found to be more functionally coherent compared to the GeMMA families generated by partitioning the GeMMA tree at any generic granularity threshold (Rentzsch and Orengo, 2013). An independent assessment by an international function prediction experiment (CAFA, see Section 3.1.3 in Chapter 3) ranked the functions predicted by assignment of sequences to DFX FunFams, among the top 10 (out of 56) function prediction methods. However, the growing

need for an improved protocol, which is unaffected by the paucity of the GO terms and annotation biases existing in the GO, necessitated the development of a new improved approach for functional classification of CATH superfamilies.

2.2 Aims and Objectives

This chapter discusses the development of a new algorithm, FunFHMMer, for capturing functional specificity and identifying functional families in CATH protein domain superfamilies using evolutionary signals in sequence alignments. This work has been published in:

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G. and Orengo, C. A. (2015). Functional classification of CATH superfamilies: a domain-based approach for protein function annotation, *Bioinformatics*, 31(21), 3460–3467.

2.3 Implementation

In this work, data from CATH (version 4.0) and Gene3D (version 12.0) was used to cluster protein domain sequences using the GeMMA clustering algorithm.

2.3.1 Development of a protocol for functional classification using sequence patterns

In order to improve the existing functional classification (DFX) in CATH, an in-depth study of the diverse TPP-dependent enzyme superfamily was carried out. This was necessary due to the following reasons - firstly, to manually analyse the performance of the existing functional classification of the superfamily using information available in the literature and functional annotation databases, and secondly, to develop strategies to use sequence patterns and other additional parameters for identification of functional families.

2.3.1.1 TPP-dependent enzyme superfamily as a preliminary test case

The Thiamine-diphosphate or Thiamine-pyrophosphate (TPP)-dependent enzyme superfamily (CATH 3.40.50.970) was chosen as a representative superfamily for the analysis of performance of the DFX protocol and test new strategies for improving functional classification. The TPP-dependent enzyme superfamily was selected as a preliminary test case since it is a very large, well-studied, functionally diverse superfamily (containing >85 unique EC terms) and the evolution of function in the TPP-dependent enzymes is very complex, as a result of gene-duplications, gene-fusions and domain-recruitment events throughout the superfamily (Duggleby, 2006; Costelloe *et al.*, 2008) (see Section 1.3.1.3 in Chapter 1 for more information about the superfamily).

Annotation of TPP-dependent superfamily domain sequences

The TPP-dependent enzyme superfamily domain sequences and their corresponding EC codes, MDA and available structural information were obtained from the Gene3D resource (Lees *et al.*, 2014). Annotations were also obtained from the Thiamine-diphosphate dependent Enzyme Engineering Database (TEED) (Widmann *et al.*, 2010) dated March 22, 2013. Gene3D domain sequences were assigned to TEED families by scanning each domain sequence with the TEED family HMMs using the HMMER3 suite of tools (Eddy, 2009). This resulted in generation of TEED-annotated Gene3D domain families for the TPP-dependent enzyme superfamily, which could now be used as a benchmark for comparing the performance of different functional classification protocols.

2.3.1.2 Exploiting specificity-determining positions in MSAs

To determine whether changes in sequence patterns could be exploited in differentiating between protein domain families, prediction of specificity-determining positions (SDPs) was carried out in FunFam alignments. A number of methods are available for SDP prediction (see Section 1.2.1.3 in Chapter 1) in multiple

sequence alignments. However, GroupSim (Capra and Singh, 2008) was chosen in this work among other SDP prediction programs because of its simple and fast implementation which could be easily integrated into the GeMMA clustering pipeline and it ranks among the best performing SDP prediction methods (Chakraborty and Chakrabarti, 2015).

Prediction of SDPs have been used in the past to generate functional subgroups for a number of selected protein superfamilies using multiple correspondence analysis (Rausell *et al.*, 2010), a phylogeny-independent stochastic approach (Mazin *et al.*, 2010) and a heuristic top-down clustering approach (Costa *et al.*, 2013). However, none of these approaches have been used for large-scale sub-classification of all known protein superfamilies. Moreover, these methods also require an accurate multiple sequence alignment of all the sequences as a starting point. As mentioned already, this can lead to erroneous sub-classification of very large or diverse superfamilies as it is difficult to obtain an accurate multiple sequence alignment of all relatives in these superfamilies (Lee *et al.*, 2010).

Prediction of SDPs in alignments using GroupSim

GroupSim (Capra and Singh, 2008) takes an input multiple sequence alignment (MSA) containing user-defined subgroups of sequences and then calculates a prediction score (G_s) for each column in the alignment except columns with more than 10% gaps overall or with a subgroup containing more than 30% gaps. G_s ranges from 0-1 where higher scores indicate a higher probability for a column in an alignment to be an SDP. However, no thresholds were defined by Capra and Singh (2008) to discriminate between conserved positions and SDPs.

To identify such a threshold for use in this present work, GroupSim was run on a SDP benchmark dataset generated by Chakraborty and Chakrabarti (2015), previously used for showing that GroupSim outperforms most other methods for SDP prediction using ranked predictions. The benchmark consisted of 20 manually-curated protein family alignments with well identified groups and SDPs

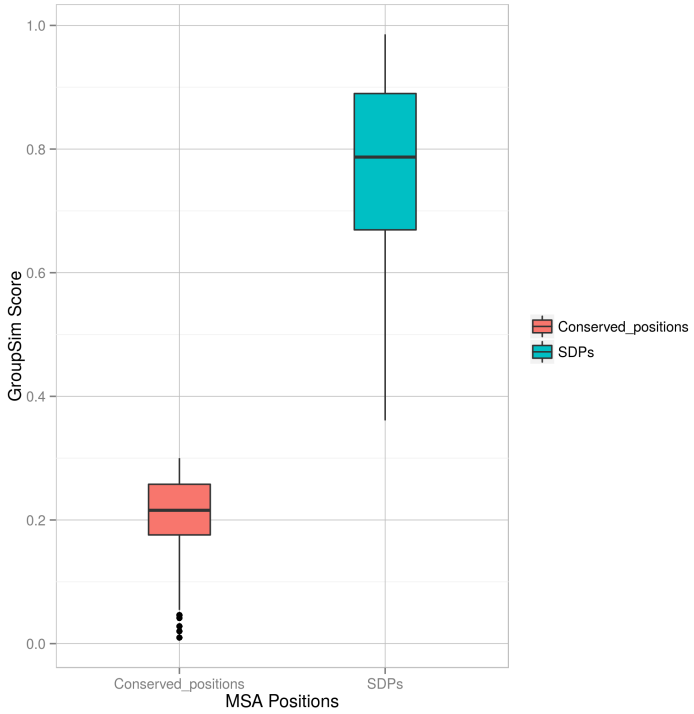


Figure 2.3: The range of GroupSim scores (G_s) for conserved positions and specificity-determining positions (SDPs) obtained for a dataset by Chakraborty and Chakrabarti (2015). Taken from Das and Orengo (2016) under CC BY 4.0.

(Chakrabarti *et al.*, 2007; Chakraborty and Chakrabarti, 2015). GroupSim was run on each alignment in the benchmark and the range of scores for conserved positions and for SDPs was determined.

In Figure 2.3 we see that the majority of conserved positions were found to have $(G_s \leq 0.3)$ and the majority of SDPs were found to be in the range $0.7 < G_s \leq 1$. Henceforth, we defined all positions with $(G_s \leq 0.3)$ as conserved positions and those with $0.7 < G_s \leq 1$ as SDPs in our subsequent analysis of multiple sequence alignments.

2.3.1.3 Prediction of SDPs in TPP-dependent enzyme families

Prediction of SDPs was carried out in all pairs of TEED-annotated Gene3D domain families having different EC annotations at the fourth level (EC4 annotations). As a control, prediction of SDPs was also performed on pairs of subfamilies within a TEED-annotated Gene3D domain family, in which all relatives have

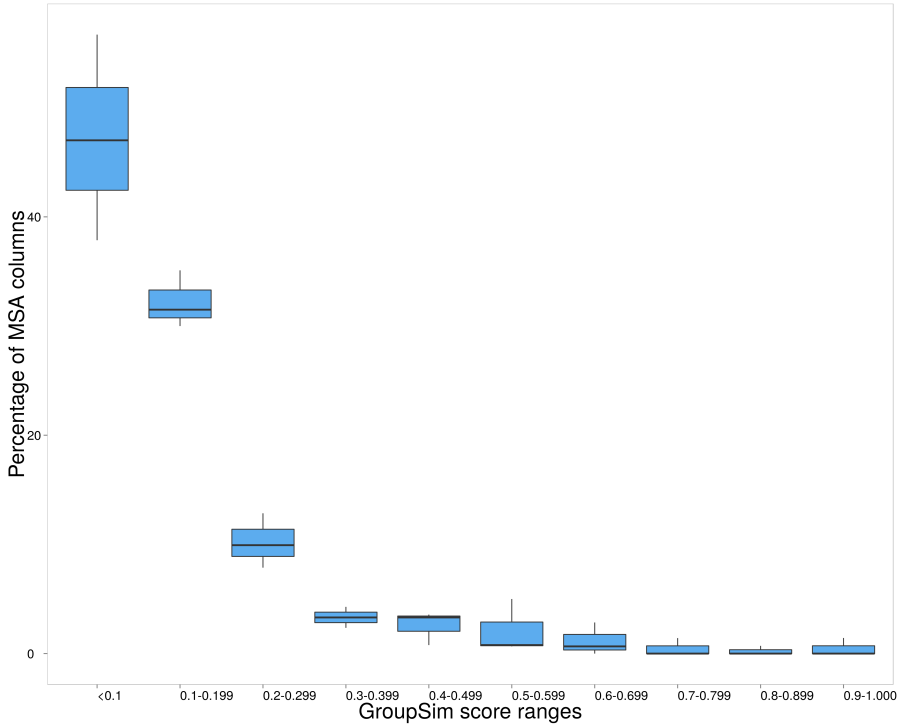
the same EC4 annotation. Only those TEED-annotated Gene3D domain families were used in this analysis that had sufficient sequence diversity for sequence analysis i.e. high information content (DOPS score $\geq$ 70, see Section 1.2.1.2 in Chapter 1).

Figure 2.4 shows that pairwise GroupSim comparisons of subfamilies belonging to the same TEED family and sharing the same EC annotation, assign the majority of residue positions as highly conserved $(G_s \leq 0.3)$ in both subfamilies and comparatively very few positions are predicted as SDPs $(0.7 < G_s \leq 1)$. In contrast, GroupSim comparisons of different TEED-annotated Gene3D families having different EC4 annotations show that although the families share a number positions that are highly conserved in both which is expected as they belong to the same superfamily, they also have a substantial number of predicted SDPs which are most likely to be implicated in the differences in the functional properties of the families.

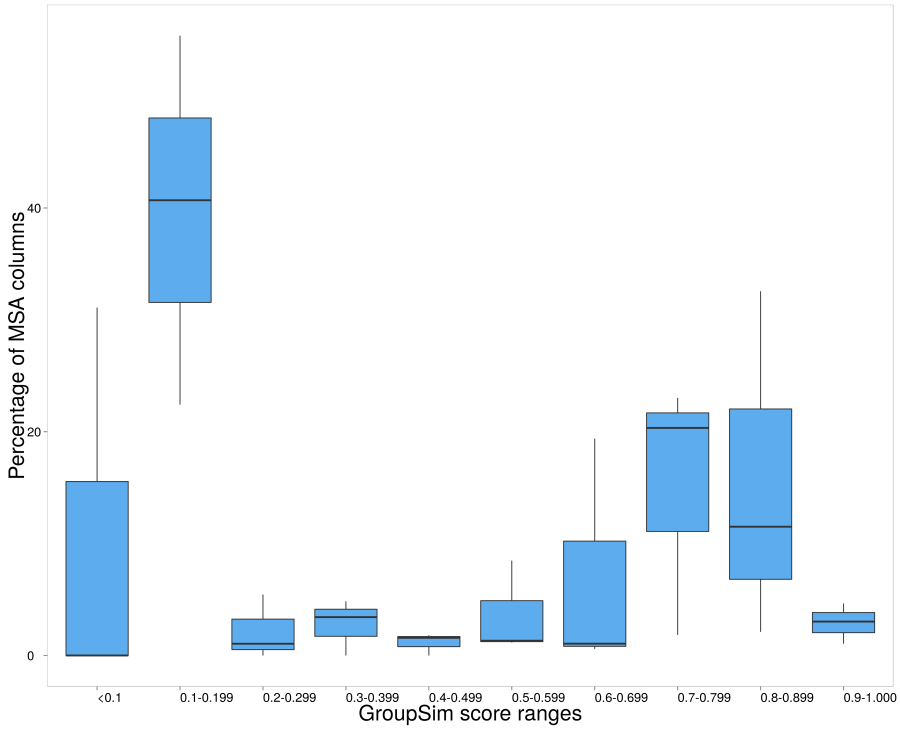
This analysis on the TPP families using GroupSim confirmed that prediction of SDPs can be exploited in analysing the functional coherence of a multiple sequence alignment. This led to the development of the family identification protocol, FunFHMMer, based on the prediction of conserved positions and SDPs.

2.3.2 FunFHMMer algorithm

The automated classification protocol, FunFHMMer, was developed to provide a functional classification method for protein superfamilies that exploits sequence patterns in order to group together sequences at the family level that share functional similarities. Similar to the DFX algorithm, the starting point of the FunFHMMer algorithm is the GeMMA clustering tree that contains only those cluster nodes that have at least one sequence annotated with high-quality GO annotations. This is done by associating each cluster node in the GeMMA tree with a set of high-quality GO annotations (see Section 2.1.4.2) from UniProt-GOA (Dimmer *et al.*, 2012), that are associated with the cluster sequences and removing any



(a)



(b)

Figure 2.4: Comparisons of **(a)** subsets of TEED-annotated Gene3D families sharing the same EC4 annotation and **(b)** different TEED-annotated Gene3D families having different EC4 annotations.

cluster which lacks at least one sequence with high-quality GO annotations.

FunFHMMer determines the optimal cut of the “bottom-up” hierarchical clustering tree generated by the GeMMA clustering algorithm (see Section 2.1.4.1) to identify functional families , also known as FunFams, in CATH protein domain superfamilies. It identifies positions that are highly conserved and SDPs in cluster multiple sequence alignments (see Figure 2.5) and calculates a novel Functional Coherence index (FC) for each parent node in the GeMMA clustering tree. This value is then used to determine whether the child nodes should be merged.

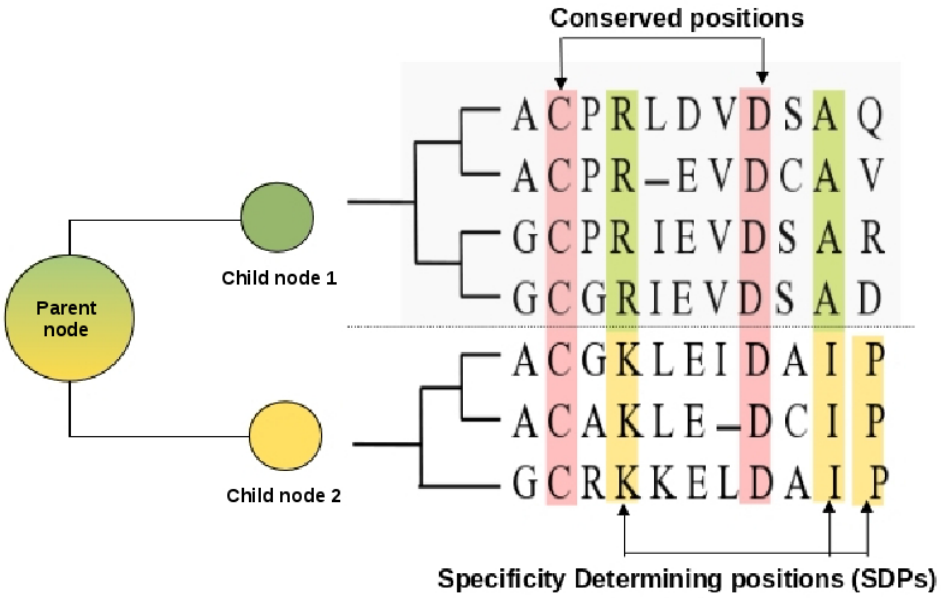


Figure 2.5: Use of predicted specificity-determining positions (SDPs) and conserved positions by FunFHMMer to infer functional coherence of cluster multiple sequence alignments (MSAs). The coloured circles represent the node sequence clusters where each colour denotes a unique function. The schematic representation of the parent node MSA and the child nodes MSA (separated by a dashed line) are shown along with the phylogenetic tree. The highly conserved positions in the MSA are shown in red and the SDPs are shown in green or yellow for different child nodes. Taken from Das and Orengo (2016) under CC BY 4.0.

2.3.2.1 Parameters affecting analysis of functional coherence of alignments

The analysis of functional coherence of a parent node multiple sequence alignment (MSA) takes into account the following parameters:

1. Information content of multiple sequence alignments. The reliability and accuracy of identifying patterns of conserved residues by sequence analysis methods rely heavily on the diversity of sequences in multiple sequence alignments (MSAs) (see 1.2.1.2 in Chapter 1). The diversity of residue conservation of positions in informative MSAs not only helps to prevent bias but also provides more discriminating conservation scores (Bartlett *et al.*, 2002a).

FunFHMMer calculates Diversity of Position Scores (DOPS) for MSAs using Scorecons (Valdar, 2002). DOPS considers the number of different conservation scores in an alignment and the relative frequency of each score, such that, DOPS is 0 if all positions in an alignment have the same conservation score and 100 when no two positions have the same conservation score. For our analysis, we have considered any alignment with a DOPS $>$ 70, as sufficiently diverse (Dessailly *et al.*, 2013). For less diverse alignments, any MSA analysis will have a higher probability of predicting false positives (false SDPs, in this case) as a result of less discriminatory conservation scores. This is because, all the sequences in less diverse alignments tend to be very similar (almost identical) to each other that results in almost identical conservation scores for all positions in the alignment. As a result, even substitutions with similar amino acids in a functionally similar homologous sequence when compared to such a less diverse alignment, would result in the prediction of SDPs for those positions.

To account for this, a DOPS factor (D_f) is used, where

$$D_f = \begin{cases} 1 & \text{if both groups have DOPS} > 70, \\ 0 & \text{if either sub-group have DOPS} < 70. \end{cases} \tag{2.7}$$

2. Proportion of predicted SDPs in a multiple sequence alignment.
FunFHMMer uses GroupSim (Capra and Singh, 2008) to predict SDPs in MSAs of parent nodes in the clustering tree. For each parent node MSA being analysed by FunFHMMer, its child nodes form the pre-defined subgroups for GroupSim. The number of SDPs (N_{sdp}) and the number of conserved positions (N_c) are cal-

culated from the GroupSim prediction scores for the parent node MSA. Whether two child nodes are merged depends on the ratio of SDPs to conserved positions in the parent node MSA (R_{sdp}). However, optimisation trials in the TPP-dependent superfamily showed that this ratio (R_{sdp}) needed to be adjusted based on the information content of alignments i.e. whether they have a low or high DOPS score.

To establish a suitable R_{sdp} ratio that ensures functional coherence for a pair of sequence clusters, we benchmarked ratios for a set of 30 large, diverse catalytic superfamilies containing at least two different EC4 annotations, in order to distinguish between parent nodes having child nodes containing sequences that share the same EC4 and those containing different EC4 annotations. Optimal strategies for calculating the R_{sdp} ratio were determined depending on whether any of the child nodes had a low DOPS score or both the child nodes had high DOPS.

For parent node alignments having child nodes that shared the different EC annotations and either of the child node had low DOPS scores, the number of predicted SDPs was found to be substantially higher than conserved positions i.e. $\frac{N_{sdp}}{N_c} > 1$. Additionally, when the child nodes shared the same EC annotations, the number of SDPs was generally found to be substantially lower than conserved positions. Thus, a simple ratio of numbers of predicted SDPs to the numbers of conserved positions could be used to distinguish between functionally coherent and functionally different parent nodes when either of its child node had a low DOPS score (Equation 2.8). Figure 2.6(a) shows that for parent nodes having any child node with low DOPS ($D_f = 0$), the $R_{sdp_{(lowDOPS)}}$ ratio (Equation 2.8) tends to be negative for parent nodes sharing the same EC4 annotation and positive when the two groups have different EC4 annotations.

$$R_{sdp_{(lowDOPS)}} = \frac{N_{sdp}}{N_c} - 1 \tag{2.8}$$

On the contrary, for parent node alignments having child nodes that had different EC annotations and both its child node had high DOPS scores, the number of

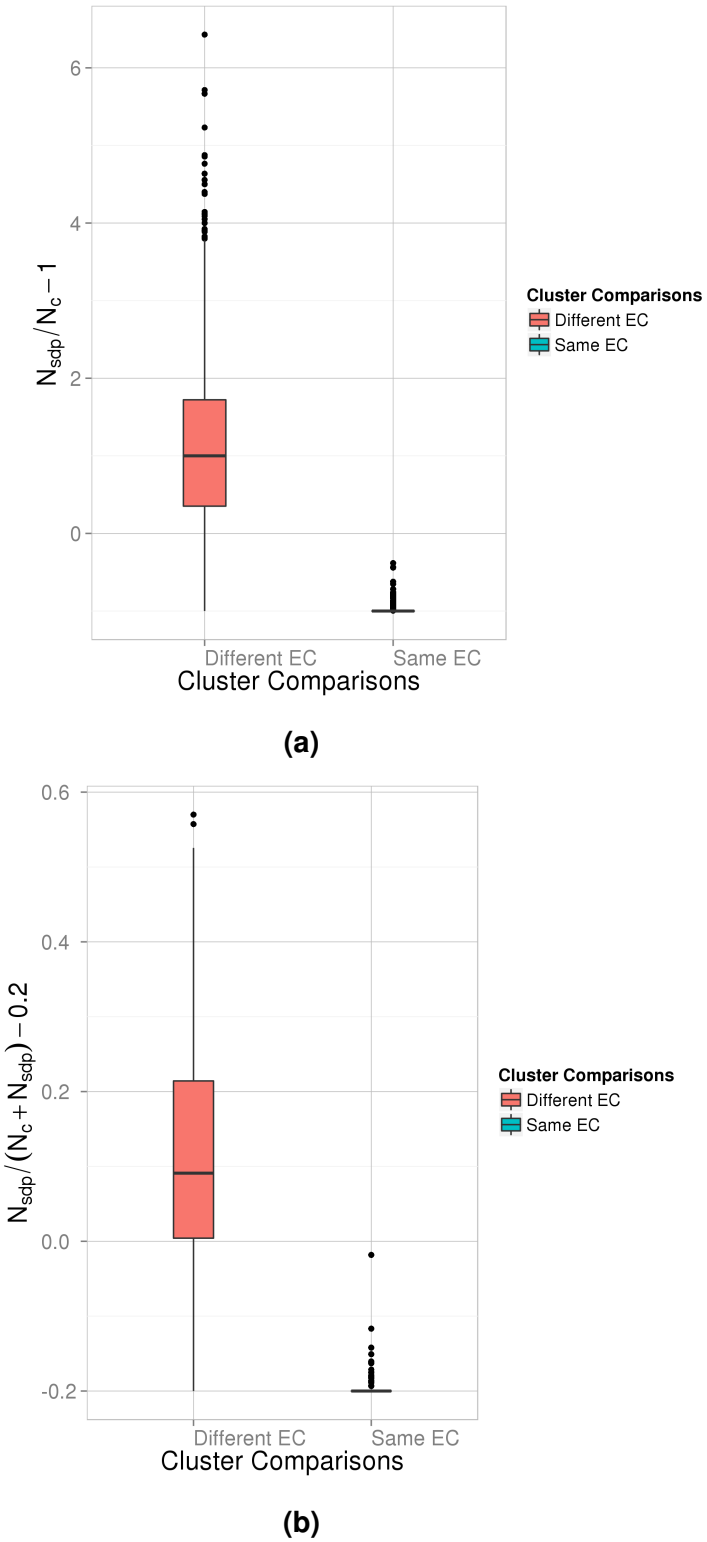


Figure 2.6: R_{sdp} ratios used to distinguish between parent nodes containing two child nodes containing the same EC annotation and those containing different EC annotations when **(a)** one or both child nodes have low DOPS, and **(b)** both child nodes have high DOPS, for 200 functionally diverse CATH enzyme superfamilies which contains at least two different EC annotations at the fourth level. Taken from Das and Orengo (2016) under CC BY 4.0.

predicted SDPs was found to be generally higher than 20% of the total number of positions in the alignment that are either conserved or a SDP i.e. $\frac{N_{sdp}}{N_c + N_{sdp}} > 0.2$. Likewise, when the child nodes had the same EC annotations, the number of predicted SDPs was found to be generally lower than 20% of the total number of positions in the alignment that are either conserved or a SDP. Thus, Equation 2.9 was used to distinguish between functionally coherent and functionally different parent nodes when both of its child nodes had high DOPS scores ($D_f = 0$). Figure 2.6(b) shows that for parent clusters with both child nodes having high DOPS scores, the $R_{sdp_{(highDOPS)}}$ ratio (Equation 2.9) also tends to be negative for parent nodes having child nodes sharing the same EC4 annotation in an MSA and positive when the two child nodes have different EC4 annotations.

$$R_{sdp_{(highDOPS)}} = \frac{N_{sdp}}{N_c \ + \ N_{sdp}} - 0.2 \tag{2.9}$$

Combining Equations 2.8 and 2.9, we get a generalized SDP Ratio (R_{sdp}):

$$R_{sdp} = \ D_f \ (\frac{N_{sdp}}{N_c \ + \ N_{sdp}} - 0.2) + (1 - D_f)(\frac{N_{sdp}}{N_c} - 1) \tag{2.10}$$

where D_f is the DOPS factor (Equation 2.7) of the MSA, N_{sdp} is the number of specificity-determining positions, N_c is the number of conserved positions in the MSA.

3. Gaps in a multiple sequence alignment. A large number of gaps in a parent node alignment would indicate that the child node alignments are of different lengths. GroupSim does not give a prediction score for columns containing more than 10% gaps overall or with a child note containing more than 30% gaps. The coherence index uses a gap factor f_{gap} which is dependant on the number of non-gapped (N_{nongap}) and gapped positions (N_{gap}) in the alignment where

$$f_{gap} = \begin{cases} 0 & \text{if } N_{nongap} > N_{gap} \text{ in an MSA,} \\ 1 & \text{if } N_{nongap} \leq N_{gap} \text{ in an MSA.} \end{cases} \tag{2.11}$$

2.3.2.2 Functional Coherence Index (FC)

The Functional Coherence index (FC) is calculated by bringing together all the above mentioned parameters using the empirical formula described below (Equation 2.12), where a coherence index of 1 indicates functional coherence of the parent node and 0 indicates that functionally diverse child nodes have been merged to form the parent node.

$$FC = \begin{cases} 1 & \text{if } R_{sdp} + f_{gap} < 0 \\ 0 & \text{if } R_{sdp} + f_{gap} \geq 0. \end{cases} \tag{2.12}$$

where, R_{sdp} is the SDP ratio (Equation 2.10) and f_{gap} is the gap factor (Equation 2.11).

The functional coherence index is used to ensure that only functionally related clusters are merged. The resulting clusters of the tree form the functional families (FunFams) for a protein domain superfamily. The workflow for the FunFHMMer algorithm is shown in Figure 2.7.

For profile-profile similarity E-values $< 10^{-50}$ between the child node alignments (calculated by COMPASS in the GeMMA clustering algorithm), the parent nodes in the GeMMA tree are assigned a coherence index (FC) of 1 since the child nodes are assumed to have significant sequence similarities and the parent nodes are functionally coherent. At higher E-values (E-value $> 10^{-50}$), the coherence index (FC) is calculated by Equation 2.12, using SDP information predicted by GroupSim. This not only helps in speeding up FunFHMMer, but also avoids prediction of false-positive SDPs which tend to arise at lower E-values because the sequences are highly similar and the DOPS values are lower.

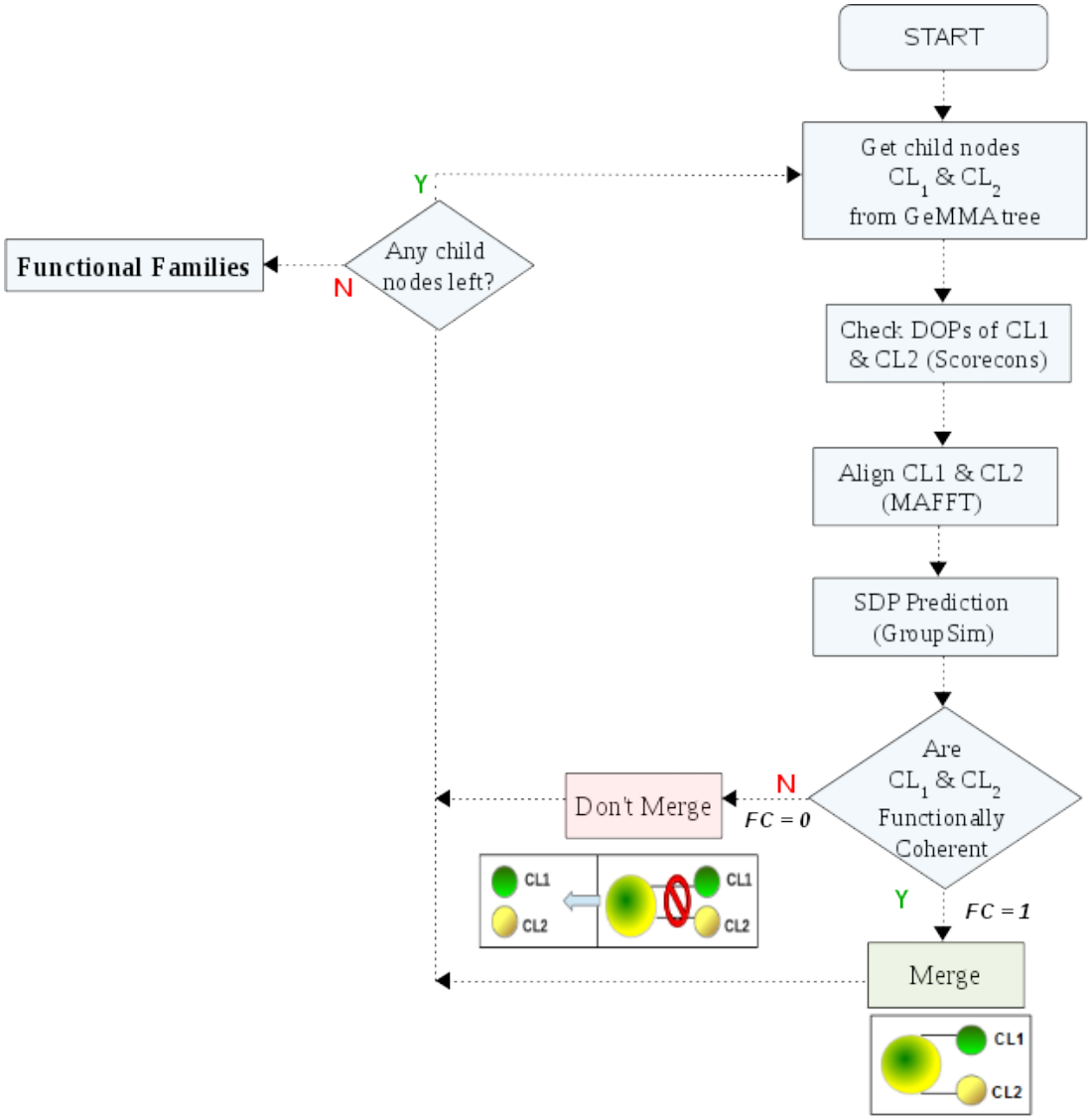


Figure 2.7: Flowchart of the FunFHMMer algorithm. FunFHMMer assesses the functional coherence (FC) of each parent node in the GeMMA clustering tree in a bottom-up manner. Taken from Das and Orengo (2016) under CC BY 4.0.

2.3.3 Modification of the GeMMA tree

FunFHMMer uses the GeMMA tree to compare the functional coherence at each parent node in a bottom-up manner to determine an optimal cut of the tree. FunFHMMer also modifies the GeMMA tree whenever analysis of a potential node merge identifies child nodes that are not functionally coherent. This is done to prevent the formation of arbitrary disjoint nodes or over-splitting of nodes. Figure 2.8 illustrates the functional family (FunFam) identification of an example superfamily by FunFHMMer in which the algorithm modifies the GeMMA superfamily tree in order to minimize the number of families generated.

While traversing the GeMMA tree, all parents nodes having an E-value $< 10^{-50}$ are created by merging its child nodes (Figure 2.8(b). For all other nodes with E-value $> 10^{-50}$, the coherence index (FC) is calculated. All parent nodes having $FC = 1$ are assumed to be functionally coherent and their child-to-parent relationship in the tree is retained. By contrast, parent nodes which are inferred to be functionally incoherent i.e. $FC = 0$ are removed and alternative routes of the tree are explored by merging the child nodes provisionally with the nearest parent nodes (Figure 2.8(c-e)). The new parent nodes resulting from this are checked for a coherence index of 1, and the modified nodes are retained if they are coherent, otherwise the child nodes are kept separate. This process is repeated up to the root of the tree following which, the leaf nodes, together with all the unmerged nodes of the tree form the FunFams of the CATH superfamily (Figure 2.8(f)).

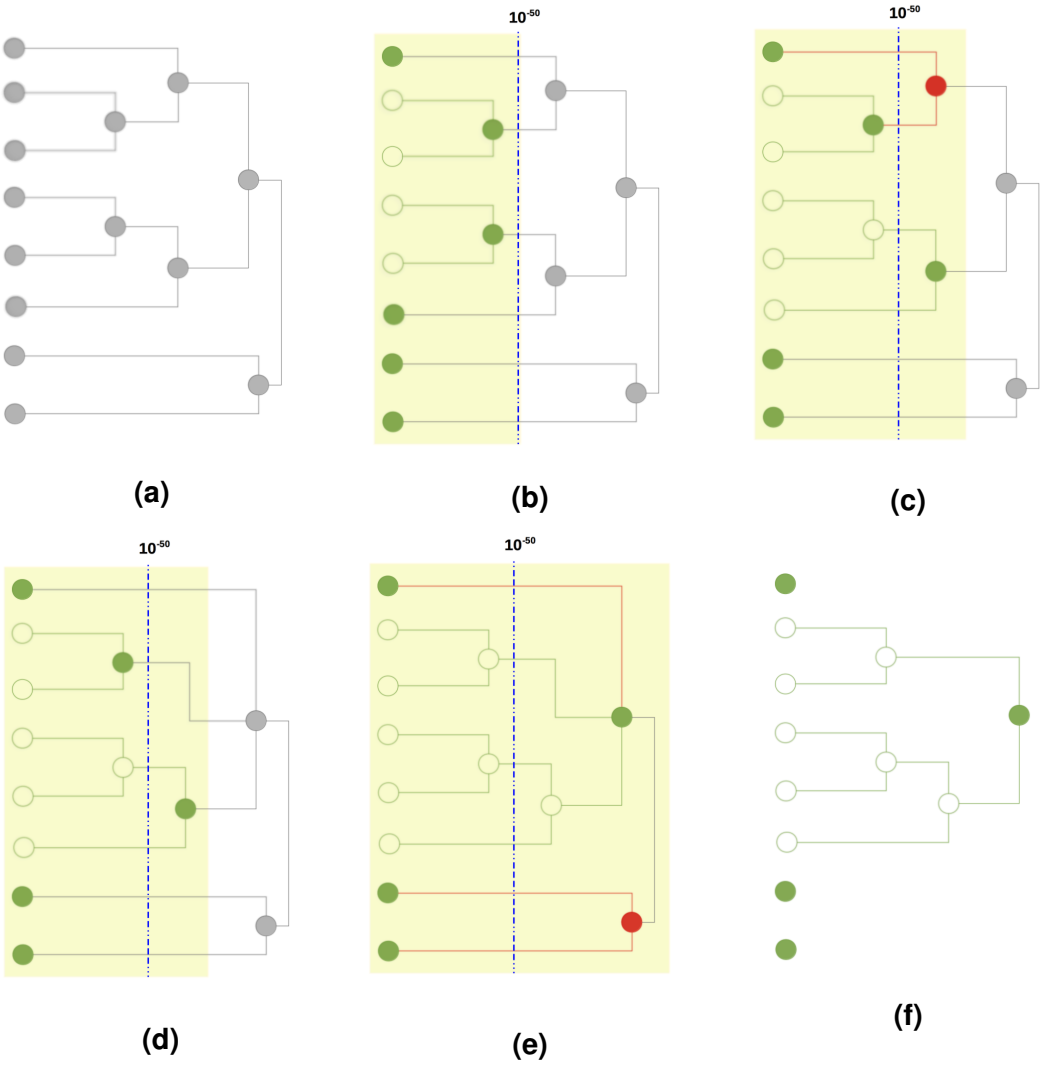


Figure 2.8: Modification of the GeMMA tree by FunFHMMer. The circles (or nodes) represent sequence clusters in the GeMMA tree. The yellow box traces the progress of FunFHMMer in processing the tree nodes and the nodes not processed by FunFHMMer are coloured grey. Functionally coherent parent nodes are coloured green and their merged child nodes are shown as unfilled circles. Clusters which are not functionally coherent are coloured red. **(a)** This shows the GeMMA tree for an example superfamily. **(b)** All nodes are merged till E-value $< 10^{-50}$ (indicated by the blue dashed line). **(c)** For E-values $> 10^{-50}$, functional coherence of nodes are calculated. For nodes that are not functionally coherent, alternative routes of the tree are explored by merging the child nodes provisionally with the nearest parent nodes. **(d)** & **(e)** The new parent nodes are checked for coherence, and the modified nodes are retained if they are coherent, otherwise the child nodes are kept separate. **(f)** When FunFHMMer finishes processing the entire tree up to the root, the leaf nodes and the outlier nodes form the FunFams of the superfamily.

2.3.4 Generation of CATH FunFams using FunFHMMer

FunFHMMer was used to generate a new set of FunFams for CATH v4.0 (Sillitoe *et al.*, 2015). A total of 110,439 FunFams were generated by FunFHMMer for 2735 CATH superfamilies. By scanning UniProtKB sequences against CATH-Gene3D and FunFam HMMs, more than 16 million sequences can be mapped to the FunFams and annotated with functional information.

FunFHMMer was also used to generate FunFams for 14,831 Pfam-A superfamilies giving 172,211 Pfam-A FunFams. This was done to assess the performance of FunFHMMer on another domain-based resource, Pfam. FunFams were also generated for the CATH superfamilies using the DFX algorithm which resulted in 26,760 DFX FunFams. This allows us to compare the performance of FunFHMMer and DFX in functionally classifying CATH superfamilies.

2.3.5 FunFam model generation and mapping of FunFam sequence and structural relatives

For each FunFam in a superfamily, an alignment is generated using MAFFT (Katoh *et al.*, 2002) and a profile hidden Markov model (HMM) is built using HMMER3 (Eddy, 2009). A model-specific inclusion threshold score is then determined for each FunFam model by choosing the lowest HMM bit score obtained by scanning all the sequences from which a model was built, against the model itself.

All sequences from Gene3D that were not clustered into an S90 cluster at the start of clustering and structural domains in the CATH superfamily are scanned against the FunFam models and a Gene3D sequence or structural domain is accepted as a new member of a FunFam if it exceeds the inclusion threshold score of the respective FunFam model.

The steps in the functional classification of CATH superfamilies are illustrated in Figure 2.9.

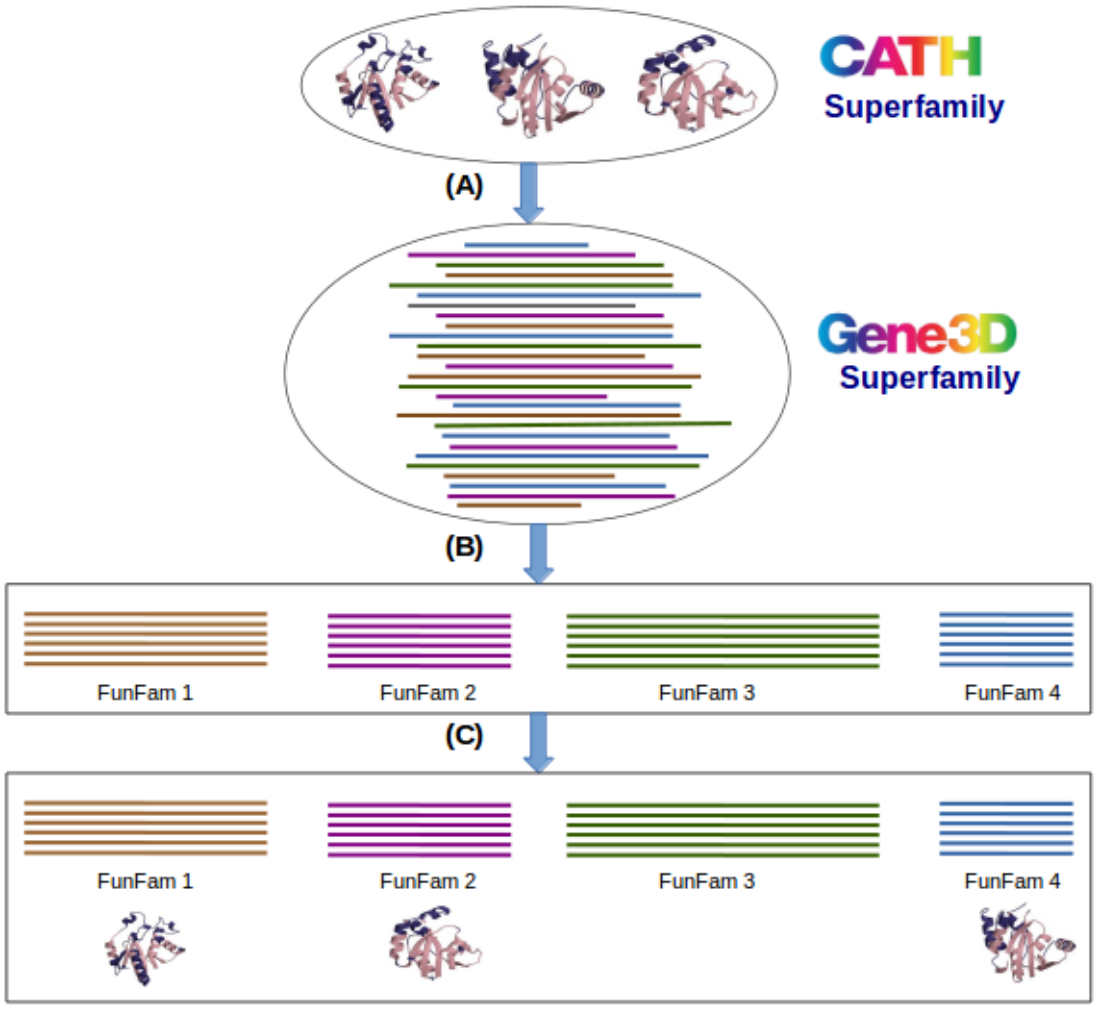


Figure 2.9: Functional classification of CATH superfamilies. **(A)** CATH superfamilies are assigned sequence relatives from UniProtKB and Ensembl in Gene3D. The colours of the sequences denote a unique function of sub-function. **(B)** Functional classification of the domain sequence relatives into FunFams (functional families) using FunFHMMer. **(C)** Assignment of CATH structural domains to the FunFams. Taken from Das and Orengo (2016).

2.3.6 Assessment of Functional Purity of FunFams

To assess whether sub-classifying the domain data in CATH-Gene3D into FunFams by FunFHMMer improved the functional purity of the FunFams, we performed the following assessments of the quality of functional classification using known functional information.

2.3.6.1 TPP-dependent enzyme superfamily

The superfamily sequences and their corresponding EC annotations were obtained from the Gene3D resource. Automated functional classification of the superfamily by FunFHMMer and DFX was benchmarked against the superfamily domain classification data obtained from TEED (Widmann *et al.*, 2010). The performance of the family identification methods on the TPP-dependent enzyme superfamily was measured using the same evaluation metrics (Performance score) as the SFLD benchmark.

The TPP-dependent superfamily is partitioned into 119 FunFHMMer families compared to 76 families by DFX. The FunFHMMer families were seen to have higher performance score (92.16) and higher purity (86.92 %) of families than DFX (Performance score =87.84, Purity = 79.69%).

Figure 2.10 shows an example from the TPP-dependent enzyme superfamily which highlights the ability of FunFHMMer to capture the functional specificity of sets of domain sequences. This feature may be useful for analysing functional shifts between FunFams. For example, conservation analysis of two FunFams consisting of sequences having the EC annotations 4.1.1.1 (Pyruvate decarboxylase) and 4.1.2.38 (Benzoin aldolase) in the TPP-dependent superfamily using Scorecons (Valdar, 2002) was performed along with identification of the SDPs between the FunFams using GroupSim (Capra and Singh, 2008). This showed that the experimentally-known catalytic site residues (extracted from the CSA (Porter *et al.*, 2004)) for domains belonging to each of these FunFams are highly conserved within each of their respective FunFam MSAs and are predicted to be

SDPs between the FunFams.

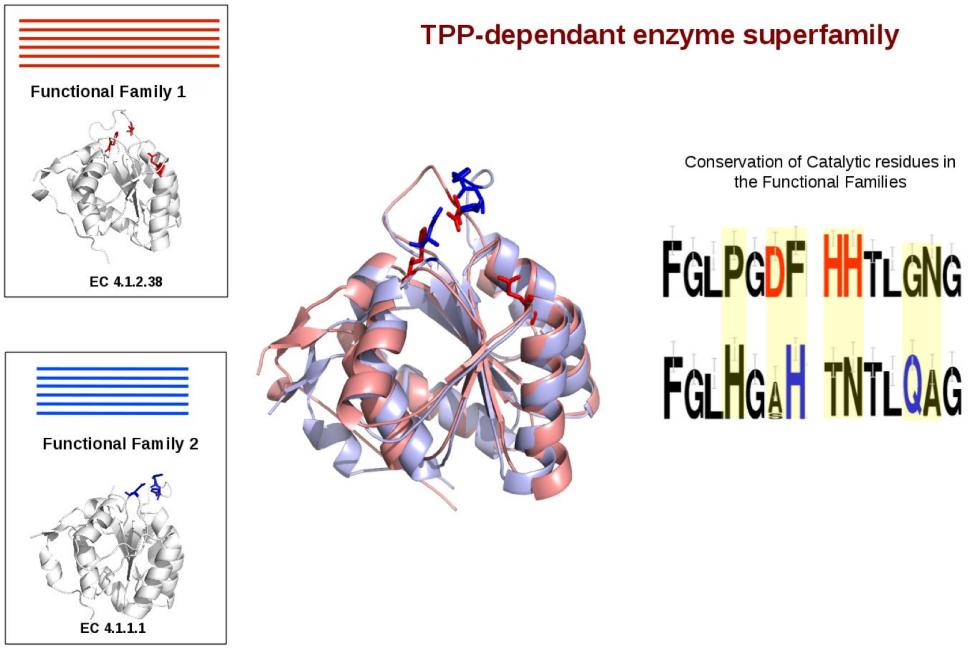


Figure 2.10: Example showing functional specificity of domains captured by FunFams generated by FunFHMMer. In this figure, two FunFams having different EC annotations (4.1.2.38 and 4.1.1.1) in the TPP-dependent enzyme superfamily are shown. The known catalytic residues belonging to domains in the FunFams 1 and 2 are shown as red and blue sticks respectively in the individual domains (shown in grey) and domain structural alignment. The catalytic residues are coloured similarly in the sequence logos of the FunFams and the SDPs between the two FunFams are highlighted in yellow. In the sequence logos, generated by WebLogo3 (Crooks *et al.*, 2004), larger residue characters indicate a greater conservation of the residues across the FunFam. Conservation analysis of the FunFams was done using using Scorecons (Valdar, 2002) and the SDP prediction was done using GroupSim (Capra and Singh, 2008).

2.3.6.2 Structure-Function Linkage Database (SFLD) superfamilies

We assessed the quality of our functional sub-classification by comparing functional assignments against the Structure-Function Linkage Database (SFLD) which has been used in benchmarking many functional classification methods for protein resources (Brown *et al.*, 2007; Lee *et al.*, 2010).

The benchmark sequences were taken from the SFLD on 24 February 2014. CATH-Gene3D domain superfamilies could be mapped onto 9 SFLD superfamilies (see Section 2.1.3.1) and a new dataset called the SFLD-Gene3D bench-

mark dataset was created comprising all the CATH-Gene3D predicted sequences mapped to SFLD whole proteins. 7 of these SFLD superfamilies comprised single domain proteins. These 7 SFLD superfamilies were mapped to single CATH superfamilies. However, the Enolase and Rubisco superfamilies contained multi-domain proteins and were each mapped to two CATH superfamilies giving a total of 11 SFLD-Gene3D superfamilies which constituted the benchmarking dataset (see Table 2.1). These superfamilies were then classified into FunFams by both DFX and FunFHMMer.

Table 2.1: Composition of the SFLD and SFLD-Gene3D benchmark dataset

SFLD superfamily	**CATH-Gene3D superfamily**	**SFLD-Gene3D mapping (%)**
Amidohydrolase	3.20.20.140	99.7
Crotonase	3.90.226.10	100
Enolase	3.20.20.120	100
Enolase	3.30.390.10	99.6
Haloacid dehalogenase	3.40.50.1000	96.7
Isoprenoid Synthase Type I	1.10.600.10	99.6
Isoprenoid Synthase Type II	1.50.10.20	100
N6P	2.120.10.30	100
Radical SAM	3.20.20.70	67.7
Rubisco	3.20.20.110	100
Rubisco	3.30.70.150	100

Performance of FunFHMMer and DFX

The performance of FunFHMMer and DFX protocol on the SFLD-Gene3D benchmark dataset can be seen in Figure 2.11. FunFHMMer outperforms DFX on the SFLD-Gene3D benchmark set on average except in the Isoprenoid Synthase Type I superfamily (CATH 1.10.600.10) where both show poor performance. The Isoprenoid Synthase Type I superfamily have a large number of sequences that have no functional annotations which makes any study relating to function in this

particular superfamily very challenging (Brown and Babbitt, 2014).

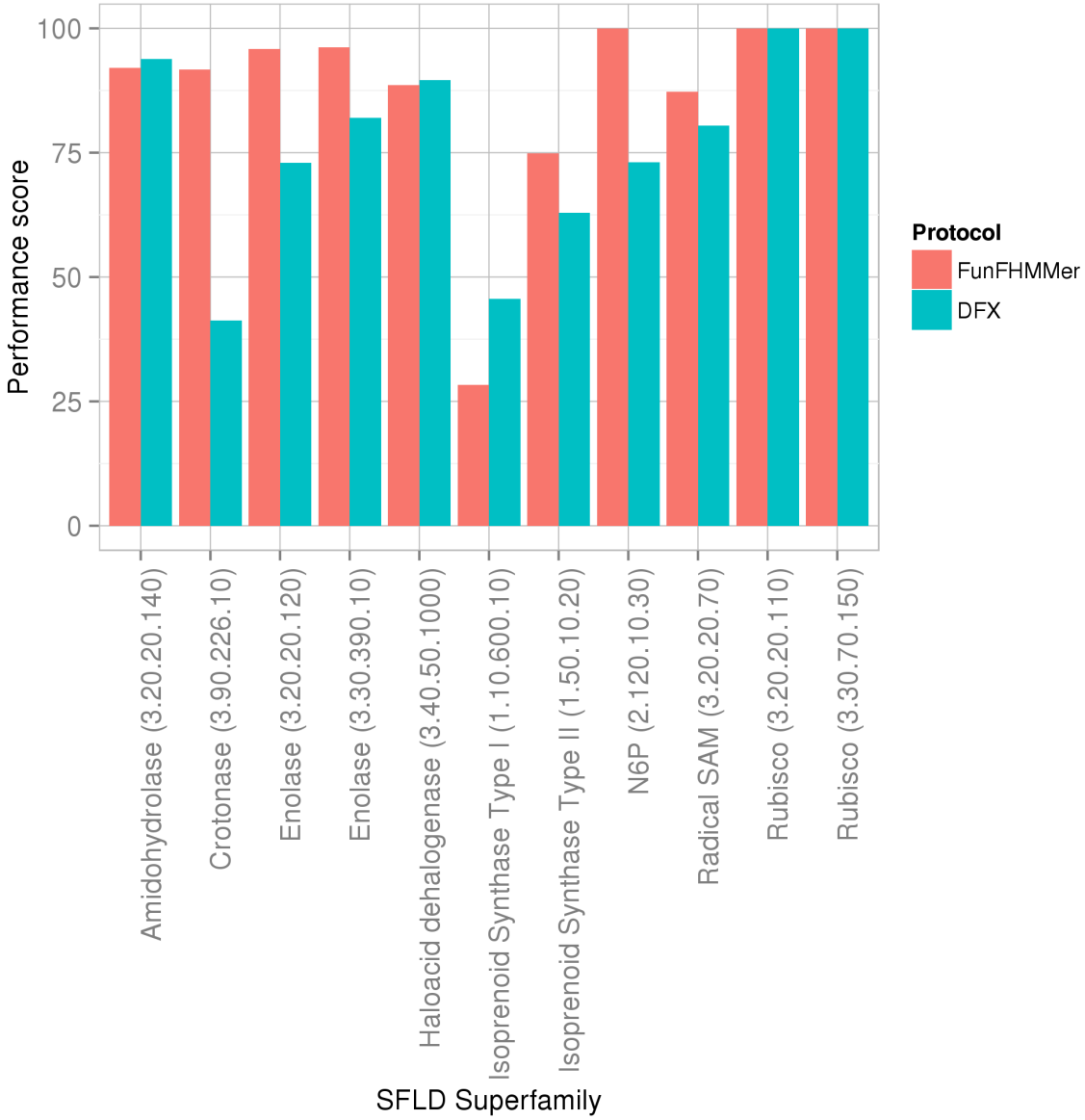


Figure 2.11: Performance of FunFHMMer and DFX on the SFLD-Gene3D benchmark dataset. Taken from Das and Orengo (2016) under CC BY 4.0.

2.3.6.3 Quality of functional classification based on EC annotations.

This test (referred to as the EC assessment hereafter) was used to analyse the performance of protein classifications in distinguishing between sequence relatives having different EC numbers. In this, we compared FunFHMMer against our previous functional classification method, DFX and other domain-family clas-

sifications i.e. Pfam and SUPERFAMILY. The domain families and superfamilies in these resources have not been explicitly classified according to enzyme function and therefore, the only purpose of including them in the assessment was to determine whether there was any benefit in function annotation transfer from sub-classification of the CATH-Gene3D resource into FunFams.

The FunFams generated by FunFHMMer for both CATH superfamilies and Pfam-A families were assessed, as were DFX FunFams, CATH superfamilies, Pfam-A families, superfamilies in SUPERFAMILY and families in SUPERFAMILY. Although CATH, Pfam and SUPERFAMILY are not publicised as functional classifications, these resources are frequently used for functional annotation of query sequences.

The EC annotations of all FunFam sequences in CATH were extracted from UniProtKB (dated February 2013) but we only considered those which had a four-digit EC number associated with the whole protein. These sequences were mapped to the different protein classifications used in the assessment and the number of different unique EC numbers per family or superfamily was analysed. The EC assessment dataset in CATH, consisting of 670,128 sequences, mapped to 1664 CATH superfamilies, 33,668 CATH FunFams generated by FunFHMMer, 9215 CATH FunFams generated by DFX, 4856 Pfam families, 24,789 Pfam FunFams generated by FunFHMMer, 1187 superfamilies in SUPERFAMILY and 2509 families in SUPERFAMILY.

Figure 2.12 shows the proportions of different sequence groupings (families or superfamilies) generated by the above-mentioned protein classifications having relatives with one or many different EC numbers. The figure has been truncated to show the proportion of families or superfamilies, up to a maximum of 10 different ECs per sequence grouping by a classification protocol. The highest proportion of families found to have only one EC number associated with them were CATH FunFams (86.5%) and the Pfam FunFams (85.5%) generated by FunFHMMer, followed by CATH FunFams (71.9%) generated by DFX, Pfam families (51.6%),

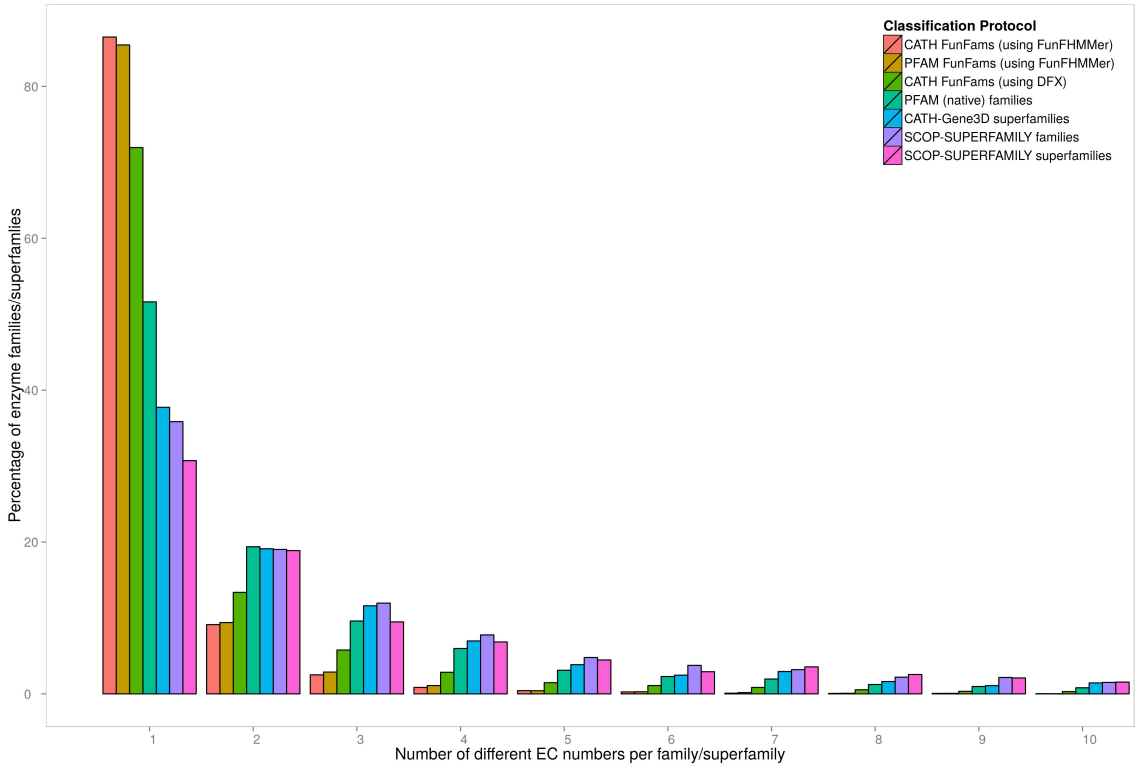


Figure 2.12: Variation of EC annotations across protein domain classifications. This figure shows the percentage of families or superfamilies having a certain number of EC annotations for each of the domain-based protein classifications. Taken from Das and Orengo (2016) under CC BY 4.0.

CATH superfamilies (37.7%), families in SUPERFAMILY (35.8%) and superfamilies in SUPERFAMILY (30.7%). This illustrates that the FunFams generated by FunFHMMer provide a more functionally coherent grouping of protein sequences than the other domain classifications. Moreover, it also shows that the FunFHMMer classification protocol is not limited in its use to CATH but can also be used to sub-classify other widely-used domain-based classification resources such as Pfam.

2.3.7 Functionally important residues highly conserved in FunFams

Ideally, FunFams are groups of protein domains with a high probability of sharing the same function(s) and therefore the functionally important residues (e.g. catalytic residues, ligand-binding residues) in a FunFam are also expected to

be highly conserved. For FunFams with sufficient information content in their MSA, residue conservation scores are calculated for each position in the alignment using Scorecons (Valdar, 2002). Scorecons scores range from 0-1 and residues having scores $\geq$ 0.7 are considered to be highly conserved (Dessailly *et al.*, 2013).

Overlaps between conserved positions in FunFams and known catalytic residues taken from the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004) were evaluated using enrichment tests adapted from Dessailly *et al.* (2013) (Figure 2.13a). For each FunFam, enrichment values were calculated as the difference between the proportion of conserved residues that are catalytic and the proportion of all residues that are catalytic. The enrichment values were averaged for each superfamily and an unpaired, one-sided Wilcoxon rank sum test (Kruskal, 1957) was run on the averaged values using the wilcox.test function in R (R-Core-Team, 2014). This test assessed a *p*-value for the null hypothesis that the proportion of catalytic residues within the conserved residues is the same as the proportion of catalytic residues within all residues in the protein domains i.e the median enrichment value is zero.

The conserved residues in all FunFam alignments were found to be significantly enriched in known catalytic residues i.e. FunFams have a greater proportion of catalytic residues within the conserved residues of a domain in comparison to all residues in the domain ($p < 3.64$ x 10^{-51}). Moreover, a comparison of enrichment scores for a subset of FunFams from 256 superfamilies, generated by FunFHMMer and DFX and sharing the same structural domains (485 domains), showed that the conserved positions in FunFams generated by both are highly enriched in catalytic residues ($p < 4.31$ x 10^{-16} for FunFHMMer and $p < 3.41$ x 10^{-15} for DFX). However, a higher proportion of FunFHMMer FunFams were found to have all known catalytic residues conserved compared to DFX (Figure 2.13b).

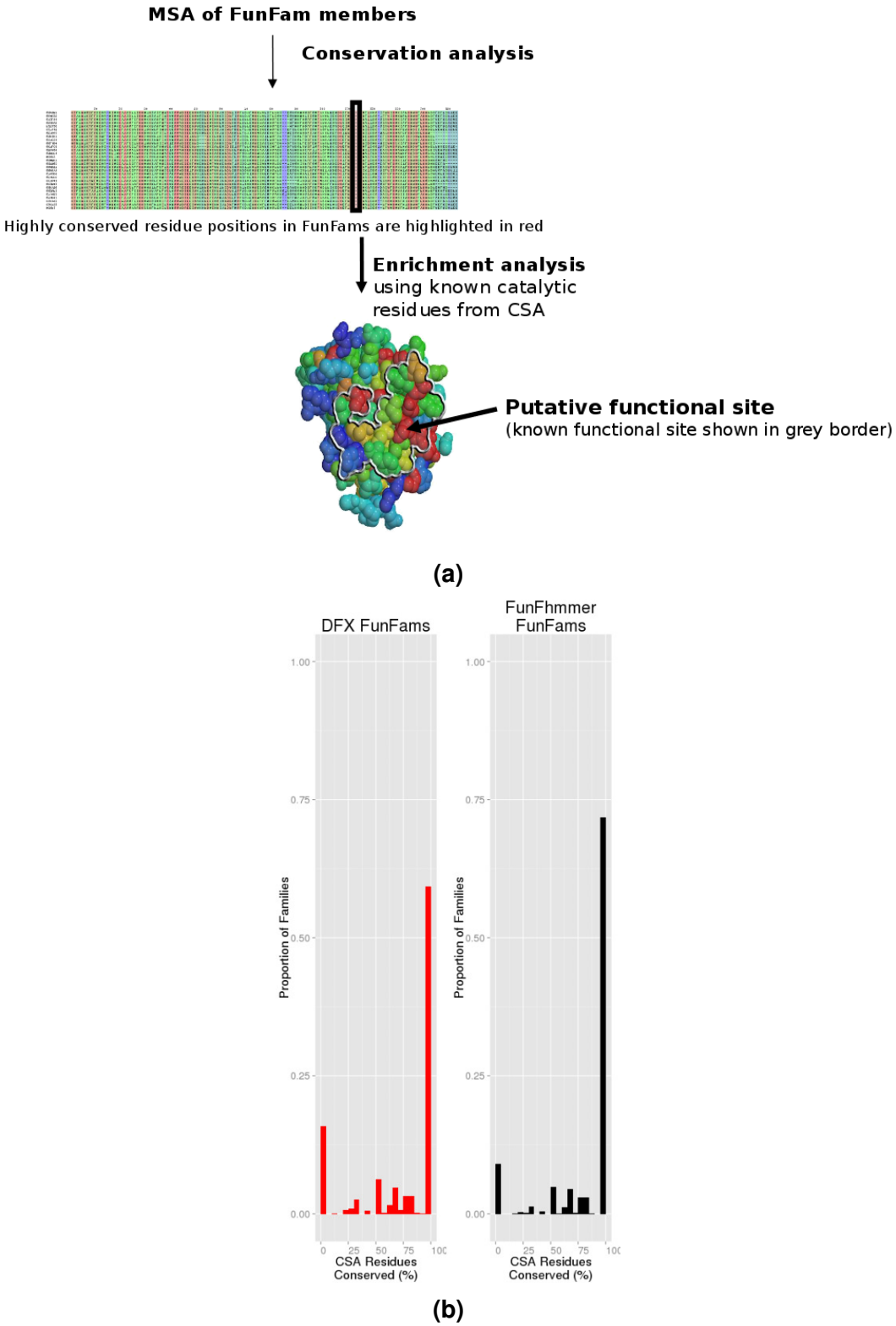


Figure 2.13: (a) Protocol for the residue enrichment analysis of FunFam alignments. **(b)** Comparison of the percentage of catalytTaken from ic residues that are conserved in FunFams generated by DFX and FunFHMMer. Taken from Das and Orengo (2016) under CC BY 4.0.

2.4 Conclusion

The FunFHMMer protocol for functional classification of CATH superfamilies was developed. The utility of such a comprehensive functional classification of protein domains is manifold – to improve our understanding of the sequence and structure mechanisms of functional divergence within a superfamily during evolution and to improve the functional annotation of uncharacterised protein domain sequences assigned to an annotated functional family within the superfamily.

An in-depth analysis of the large, well-studied and diverse Thiamine pyrophosphate (TPP)-dependent enzyme superfamily was first performed to determine sequence-based parameters that are critical for inferring functional coherence of sequence alignments. These parameters were then incorporated in the FunFHMMer protocol to calculate a novel index to assess functional coherence of sequence alignments.

The FunFHMMer functional classification protocol was used to functionally classify 2735 protein domain superfamilies in CATH-Gene3D that results in 110,439 functional families or FunFams. It was found to be able to separate the FunFams exploiting residue conservation and differences in specificity-determining positions (SDPs). In this chapter, it is highlighted that the FunFHMMer protocol results in FunFams that are significantly more functionally pure than the previous classification protocol in CATH, DFX, reported in 2013 (Sillitoe *et al.*, 2013). This was demonstrated using three independent benchmarking protocols based on the manually-curated SFLD superfamilies, consistency of the EC annotations (Bairoch, 2000) within the superfamilies and an analysis on conservation of known functional sites in FunFams. The FunFams have been found to be structurally coherent by Garcia *et al.* (2016) which indicates that they may be a good resource for searching templates for homology modelling. Furthermore, the utility of FunFams in annotating metagenome data (Dawson, 2015) and identifying new drug targets (Garcia *et al.*, 2016) was also demonstrated recently.

All FunFam data are made available through the CATH web-pages (`http:`

//www.cathdb.info/) which provides a listing of FunFams within each superfamily. For each FunFam, visualization of the multiple-sequence alignment (also available for download) and information regarding functional annotations (i.e. EC and GO annotations), the multi-domain architectures and taxonomy of the sequence relatives are provided.

Chapter 3

Protein function annotation using FunFHMMer

3.1 Background

The Genomes Online Database (Reddy *et al.*, 2014), which is a centralized resource of genome-sequencing projects worldwide, lists $>$ 64,000 sequencing projects as of June 2015, and these are expected to hugely increase the numbers of known sequences in UniProtKB. In contrast, $<$ 1% of the protein annotations in the current UniProtKB database are experimentally validated. Since the current rate of experimental annotations and manual curation process will never be sufficient for complete annotation of the proteins captured in public databases (Baumgartner *et al.*, 2007), the gap between uncharacterised sequences and annotations will continue to rise. In order to bridge this gap, computational function prediction and annotation approaches will be essential.

3.1.1 Current approaches for protein function prediction

3.1.1.1 Sequence homology

The conventional method used for protein function annotation is a sequence homology search followed by annotation transfer, based on the principle that evolutionarily-related proteins having high sequence similarity have similar, if not identical functions. Several studies have investigated the accuracy of directly inheriting functional annotations for uncharacterised sequences from a homologue having known functions.

Initially, three studies (Devos and Valencia, 2000; Wilson *et al.*, 2000; Todd *et al.*, 2001) had suggested that enzyme function is generally conserved at sequence identities above either 40% (Wilson *et al.*, 2000; Todd *et al.*, 2001) or

50% (Devos and Valencia, 2000). These three studies performed all-against-all pairwise sequence comparisons of proteins with known structures in their dataset and examined the similarity in EC numbers at different sequence identity thresholds, however, their analysis did not distinguish between single and multi-domain proteins. Soon after, Hegyi and Gerstein (2001) reported that multi-domain proteins have significantly less functional conservation (approximately two-fold) than single-domain proteins unless they share the same multi-domain architecture. Rost (2002) contested the sequence identity thresholds for enzyme function conservation suggested by earlier studies (Devos and Valencia, 2000; Wilson *et al.*, 2000; Todd *et al.*, 2001) and argued that the datasets used in the previous studies were either small or did not account for the compositional biases existing in the databases. Rost (2002) accounted for database biasses in his own analysis by grouping protein sequences into families based on sequence similarity and selecting representative sequences from each family to construct an unbiased dataset. The conservation of enzyme function within sequences in the unbiased dataset were then compared with those from the original biased dataset. The biased dataset showed results similar to that of the previous studies i.e. the entire EC number was conserved above $\sim$ 50%. In contrast, using the unbiased dataset it was seen that both the first and all four EC numbers start diverging below 70% sequence identity. Soon after, Tian and Skolnick (2003) analysed enzyme function conservation, in a manner similar to Rost's, by classifying enzyme families based on both sequence identity and functional similarity i.e. sequences sharing the same four digit or the same first three EC numbers. Their analysis suggested that sequence identity of above 60% is required to inherit entire EC numbers with at least 90% accuracy. Addou *et al.* (2009) revisited the analyses of the Rost and Skolnick groups a few years later and reported that sequences having homologs with 40% and 60% pairwise sequence identity were still sufficient to safely inherit the first three and entire EC numbers, respectively. Furthermore, for multi-domain proteins, the pairwise sequence identity thresholds increase to 50% and 70% for

the first three and entire EC numbers respectively on the domain level, for safe inheritance of EC numbers (Addou *et al.*, 2009).

A typical sequence homology-based prediction method involves a homology search using the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) to identify similar sequences from a sequence database, followed by functional annotation transfer. PSI-BLAST (Altschul *et al.*, 1997) is often used instead of BLAST as it is much more sensitive since it performs profile-to-sequence comparisons rather than sequence-to-sequence comparisons.

3.1.1.2 Protein family resources

Protein family resources cluster protein sequences into families and subfamilies based on their sequence, structure or functional similarity (in the case of annotated protein sequences). These family resources may be used for annotating uncharacterised sequences by mapping query sequences to the best matched family and inheriting the annotations from the characterised sequences.

Manually-curated GO term associations are readily available from certain family resources such as TIGRFAM (TIGRFAM2GO) and HAMAP (HAMAP2GO). BAR+ (Piovesan *et al.*, 2011), an automated annotation method based on the annotation transfer from protein families, produces clusters such that the pairwise sequence identity between relatives in a cluster is 40% with at least 90% of sequences in the pairwise alignment overlapping. A BLAST search of query sequences against the BAR+ clusters is performed and statistically validated GO and Pfam annotations are then inferred for the sequences based on the sequence identity and coverage of the matches (Piovesan *et al.*, 2013).

A domain-centric approach can also be exploited in functional annotation of the whole protein by identifying domains within a sequence, associating functions to these domains from a domain-based family resource e.g. Pfam (Finn *et al.*, 2014) or CATH (Sillitoe *et al.*, 2015) and integrating these functions in order to describe the function of the whole protein. Manually-curated GO associa-

tions for protein domain families are available for ProDom (ProDom2GO), Pfam (Pfam2GO) and InterPro (InterPro2GO) (Camon *et al.*, 2004). Various automated methods have been developed in recent years to exploit the functional signal encoded in domains to annotate uncharacterised proteins (Table 3.1).

Table 3.1: Protein function annotation methods based on protein domain families.

Prediction Method	**Domain Resource**	**References**
GO predictions from ProDom and CDD	ProDom and CDD	(Schug *et al.*, 2002)
GOtrees	Pfam	(Hayete and Bienkowska, 2005)
MultiPfam2GO & Naïve Bayesian model	Pfam	(Forslund and Sonnhammer, 2008)
SCOP2GO	SCOP	(Lopez and Pazos, 2013)
dcGO	SCOP, SUPERFAMILY	(Fang and Gough, 2013)
DFX	CATH-Gene3D	(Rentzsch and Orengo, 2013)
FunFHMMer	CATH-Gene3D	(Das *et al.*, 2015b,?)

Schug and co-workers (Schug *et al.*, 2002) developed a rule-based association of GO terms to ProDom (Bru *et al.*, 2005) and CDD (Marchler-Bauer *et al.*, 2014) domains for which thresholds were also determined. Query sequences were annotated by performing a BLAST search against ProDom or CDD followed by annotation transfer from matched domains that met the thresholds of domain-function associations. The GOtrees method (Hayete and Bienkowska, 2005) used decision trees to predict GO terms for query sequences based on domain composition in proteins (from Pfam) and other sequence features. Forslund and Sonnhammer (Forslund and Sonnhammer, 2008) extended the Pfam2GO approach and developed two protocols: a rule-based (MultiPfam2GO) model that assigns a GO term to a domain if all proteins containing the domain are annotated with that GO term and a naïve Bayesian model, which associates GO terms to domains probabilistically. The SCOP2GO (Lopez and Pazos, 2013) method associates MFO terms to SCOP structural domains and annotates query sequences by scanning them against PSSM libraries that are built for SCOP domains having

same fold and function (i.e. same GO terms). dcGO (Fang and Gough, 2013, domain-centric GO) predictor infers GO terms for individual SCOP domains or supradomains (two or more domains which are known to function together) based on whole protein annotations from UniProtKB-GOA and domain architecture information extracted from SUPERFAMILY.

DFX (Rentzsch and Orengo, 2013)(described in Section 2.1.4.2 in Chapter 2) classifies the protein domain superfamilies in the CATH-Gene3D resource into domain functional families or FunFams using GO-based cluster evaluation of the hierarchical clustering algorithm, GeMMA (described in Section 2.1.4.1 in Chapter 2). Each FunFam is associated with GO terms probabilistically based on GO annotations of parent proteins of its domain sequences, which are then used to annotate query sequences based on their CATH domain composition. FunFHMMer (Das *et al.*, 2015b) (described in Section 2.3.2 in Chapter 2) is an improved method for functional classification of CATH-Gene3D superfamilies which evaluates functional coherence of clusters using the evolutionary signals in cluster alignments and outperforms DFX and other domain-based classification protocols in predicting protein function.

3.1.1.3 Gene Ontology-based prediction methods

Gene Ontology-based prediction methods first use sequence comparison methods such as BLAST or PSI-BLAST to identify sequence homologs with known GO annotations. The GO annotations from the homologs are then utilised in different ways by different GO-based function prediction methods. Some GO-based methods like GOtcha and PFP (Hawkins *et al.*, 2006, 2009) predict GO terms by combining the GO annotations of the homologs. For example, the PFP method (Hawkins *et al.*, 2009) predicts the function of a query sequence by combining the frequency of GO terms of a wide range of E-value (up to E-values of 100) sequence matches of a PSI-BLAST search for the query sequence using an E-value based scoring scheme along with a data-mining tool, Function Association

Matrix, that predicts additional GO terms for the sequence hits from PSI-BLAST based on the frequency at which they co-occur in UniProt sequences (Hawkins *et al.*, 2006).

In contrast, methods like ConFunc (Wass and Sternberg, 2008) and GoFDR (Gong *et al.*, 2016) sub-group PSI-BLAST sequence hits for a query sequence according to their GO annotations such that for each GO term, the PSI-BLAST homologs sharing the target GO term annotation are grouped together to form a sub-group. ConFunc (Wass and Sternberg, 2008) calculates residue conservation scores for each sub-group alignment to identify conserved residues and generates position-specific scoring matrix (PSSM) profiles for the sub-group alignments which are scored against the query sequence to predict functions. ConFunc uses the PSSM profile *E*-value scores along with the frequency of the GO terms within the PSI-BLAST search hits to provide the confidence scores for the GO term predictions. In contrary, GoFDR identifies, for each GO term, functionally discriminating residues or FDRs (referred to as specificity-determining residues or SDPs in this work; see Section 1.2.1.3 in Chapter 1) between an alignment of PSI-BLAST homologs sharing the target GO term annotation (termed as homo-functional MSA) and an alignment of homologs lacking the target GO term annotation (termed as hetero-functional MSA). GoFDR then builds a PSSM for the FDRs and scores the query sequence for its association with the target GO term. Finally, it converts the PSSM profile scores into probabilities by using a conversion table created by training GoFDR with a large number of sequences.

3.1.1.4 Phylocogenomics

Phylogenomics-based (Eisen, 1998) function annotation methods are based on the principle that in certain cases of sequence homology, the most similar sequences will not always correspond to similarity in function as homologous sequences can be orthologous or paralogous. Phylogenomics-based annotation

methods make use of the evolutionary history of putative homologs of the query sequence followed by function transfer from the closest ortholog.

SIFTER (Statistical Inference of Function Through Evolutionary Relationships) (Engelhardt *et al.*, 2006) is a statistical graphical model for predicting protein molecular functions using phylogenomics. It checks for the closest ortholog to the query by segregating orthologous and paralogous events of the related gene and inferring gene duplications on a gene tree and comparing it with a species tree. It then transfers the available functional annotations of the ortholog to the query sequence.

3.1.1.5 Structural homology

Knowledge of protein structure plays an important role in protein function prediction since protein structures are conserved even in the absence of any sequence similarity. Such distant evolutionary relationships can be captured by using protein structure comparison methods like SSAP (Taylor and Orengo, 1989), CE (Shindyalov and Bourne, 1998) and DALI (Holm and Sander, 1995). These methods use the Protein Data Bank or structure classification databases, CATH (Orengo *et al.*, 1997) and SCOP (Murzin *et al.*, 1995) as the source of protein structure relatives. In cases of high structural similarity, functional similarity can be suggested, however, Martin *et al.* (1998) showed that protein fold similarity may not be always sufficient to conclude functional similarity as many proteins having the same function can have different folds and vice-versa.

Various algorithms make use of other protein structural data like 3-dimensional (3D) patterns, pockets or clefts that help in function annotation. Protein surface (binding pockets and clefts) prediction methods also provide useful information about likely protein functional sites using methods like pvSOAR (Binkowski *et al.*, 2004), CASTp (Dundas *et al.*, 2006), SiteEngine (Shulman-Peleg *et al.*, 2005) and THEMATICS (Ondrechen *et al.*, 2001).

3.1.1.6 Combination of heterogenous data

The protein sequence-structure-function relationship is very complex and a similarity in either protein sequence or structure does not always imply functional similarity. Sequence or structure homology-based function prediction methods can often lead to erroneous functional assignments (Devos and Valencia, 2000; Punta and Ofran, 2008) which may arise due to annotations transferred from paralogs (Theißen, 2002) or from proteins within the twilight zone of similarities i.e. $<30\%$ sequence identity (Chung and Subbiah, 1996), multi-functional proteins (Jeffery, 2003), domain-shuffling in multi-domain proteins (Bashton and Chothia, 2007) or due to misannotations existing in the databases (Devos and Valencia, 2001).

Proteins can acquire new functions from a combination of mechanisms such as gene duplication, gene fusion, gene recruitment, oligomerisation, alternative splicing and post-translational modifications (Todd *et al.*, 2001). As a result, a large number of protein function prediction methods combine data from heterogeneous sources in order to predict functions of uncharacterised proteins since most targets are hard to characterise using a single method. In such cases, when the predictions of several methods show consensus or indicate a similar function for the protein, there is greater confidence in the predictions. Many protein function prediction methods are available as web servers such as ProFunc (Laskowski *et al.*, 2003), ProKnow (Pal and Eisenberg, 2005) and PredictProtein (Yachdav *et al.*, 2014) and combine several sequence-based and structure-based methods to predict functions.

Recently, an increasing number of methods utilise machine-learning to combine data from different methods or sources to predict GO terms (Clark and Radivojac, 2011; Wass *et al.*, 2012; Cozzetto *et al.*, 2013). For example, CombFunc (Wass *et al.*, 2012) first predicts GO terms using different methods separately and then the features for GO terms identified by each method are combined using a support vector machine (SVM) to make the final predictions. The individual meth-

ods that CombFunc uses include sequence homology using BLAST/PSI-BLAST, domain-based predictions using information from InterPro and predictions from protein-protein interaction and gene expression data. FunctionSpace (Cozzetto *et al.*, 2013) is a another recent machine-learning method which combines information from a wide variety of sources. It integrates sequence, gene expression, protein-protein interaction data and UniProtKB annotations retrieved by a text-mining tool into a single framework. Information from all these methods are then combined in a probabilistic manner taking into account the ontology structure of GO to predict GO terms for a query sequence.

3.1.2 Assessment of function prediction methods

A large number of function annotations are available today which provide computational function annotations at the protein level exploiting different approaches. However, it is essential for both computational and experimental biologists to know the accuracy of these methods in order to understand which prediction approach performs better and to decide whether the automated function annotations provided by the methods can be relied upon. Moreover, this would help us in understanding the strengths and weaknesses of different approaches of function predictions.

There are two major challenges associated with comparative assessment of automated function prediction programs (Godzik *et al.*, 2007). The first being the requirement of an accurately annotated target benchmark dataset for assessment. An ideal benchmark dataset will constitute unannotated proteins that do not show significant sequence or structural similarity to annotated proteins so that these can be used to test the function inference of the methods or how well they make a well-informed guess. Furthermore, the benchmark dataset must be unbiased such that it contains sequences from all kingdoms across the tree of life. One approach of building a benchmark dataset is to use those protein sequences whose functions have been recently determined but have not yet been published

in databases or other resources. Another approach involves a rollback dataset in which data from a database is used by the function prediction method up to a particular date e.g till June 2013 and then predictions are made for sequences which have accumulated after June 2013.

The second challenge is the establishment of an evaluation metric for accurate assessment of the performance of automated function prediction methods. For example, when two function prediction methods are compared on the basis of their ability to predict the function of a protein and both fail to accurately predict the known protein function, it is necessary for the assessment metric to differentiate between a near-miss or wide-miss by quantitating the differences between the predicted functions and the true function (Godzik *et al.*, 2007). This also reiterates the importance of using a suitable system for describing protein functions which facilitates computation and an associated evaluation metric that can be used to quantitate distance between its functional terms. There have been many attempts at assessments of automated function prediction so far, which have been summarized in Figure 3.1.

AFP Assessment Experiments	Target Benchmark Dataset	Particip-ating Methods/ Groups	Evaluation method	Challenges	Conclusions
GASP , 1999 (Genome annotation assessment in *Drosophila melanogaster*) (Reese *et al.*, 2000)	*Adh* region (2.9 Mb) of *Drosophila melanogaster* genome	12 groups	Specificity and Sensitivity evaluated on the basis of unreleased cDNA sequences and experimental annotations.	Lack of an absolutely correct standard against which predictions can be evaluated.	Gene predictions are useful for genome-scale annotations.
CASP6, 2004 and **CASP7, 2007** (Critical Assessment of Techniques for Protein Structure Prediction) (Calace *et al.*, 2006)	Function prediction of proteins used in the CASP structure prediction category.	23 predictions from 18 research groups	Comparative assessment of function predictions among different methods.	Experimental function annotation of the targets was not available at the time of assessment .	i) Higher reliability can be assigned to cases where completely independent methods give the same or similar predictions. ii) Predicting functional sites should be established as another category.
AFP, 2005 (First Automated Function Prediction Meeting) (Godzik *et al.*, 2007)	5 proteins that were not homologous to any annotated proteins at that time.	8 prediction methods	GO based Semantic Similarity measure.	Development of a viable, blind benchmark.	Collaboration between computational and experimental scientists is required to provide challenging cases for function prediction programs.
MouseFunc, 2006 (Pena-Castillo *et al.*, 2008)	A training set of genes along with gene properties and biological relationships among them was provided along with a set of genes for function prediction.	9 groups	GO term predictions were evaluated by Precision-Recall and ROC (receiver operating characteristic) curves.	Evaluation of machine-learning methods	All methods performed similarly. Molecular function terms were easier to predict than for biological process.
CAFA , 2010-2011 (Critical Assessment of protein Function Annotation algorithms) (Radivojac *et al.*, 2013)	48,298 Swiss-Prot protein sequences from 7 eukaryotic and 11 prokaryotic species were released. Methods were evaluated on 866 proteins from 7 eukaryotic and 4 prokaryotic species, which accumulated annotations.	54 methods from 23 research groups	GO term predictions were analyzed by: i) Protein-centric metric - F_{max} from Precision-Recall curves. ii) Term-centric metric - Specificity and Sensitivity from ROC curve.	i) Incomplete experimental annotations and biases involved. ii) Complexities involved in describing function using GO. iii) Proteins can be multi-functional and promiscuous	i) Top methods outperform BLAST. ii) Considerable need for improvement of currently available tools.

Figure 3.1: Tabular representation of assessment experiments of automated function annotation methods.

3.1.3 Critical Assessment of Function Annotation (CAFA)

The Critical Assessment of Function Annotation (CAFA) experiment is a recent major bioinformatics initiative conducted by the Automated Function Prediction Special Interest Group (AFP-SIG), which aims to provide large-scale assessment of computational function prediction algorithms using a time challenge (Friedberg and Radivojac, 2016; Radivojac *et al.*, 2013). A dataset of proteins lacking any experimental GO terms are selected as targets by the CAFA organisers and are provided to the automated function prediction community six months before the submission deadline. During this time, the target sequences are annotated by assignment of GO terms along with confidence scores by function prediction algorithms and the predictions are submitted to the CAFA organisers. The set of all experimentally annotated proteins available on the submission deadline, forms the training dataset. After the submission deadline, the experimental annotations are allowed to accumulate over a period of approximately 12 months. The function prediction methods are then evaluated on the subset of target proteins (benchmark dataset) that accumulate experimental GO terms during the accumulation phase.

The performance of all the participating methods are compared to two baseline methods: (i) BLAST, based on the search results using BLAST software against the training database; and (ii) a Naïve method, which predicts all GO terms in a GO ontology for each target with the relative frequency of the GO term in the training database over all annotated proteins (Jiang *et al.*, 2016; Radivojac *et al.*, 2013).

3.1.3.1 CAFA evaluation metrics

The performance accuracy of protein function prediction methods that predict all GO terms associated with a particular protein sequence are evaluated using the following metrics:

(i) Precision-recall ($pr - rc$) curves

For each target and some decision threshold $\tau \in$ [0,1], the GO terms in an ontology are assigned to it with confidence scores greater than or equal to τ, were propagated up the GO hierarchy or directed acyclic graph (DAG) to the root, yielding the set of predicted GO terms for that target (predicted set). The true (experimental) GO terms are also up-propagated the GO hierarchy for every target (true set). Any terms which overlap between the predicted and the true set were considered as correct at that decision threshold τ (see Figure 3.2). Consequently, the precision (pr) and recall (rc) for each target were computed as:

$$pr_i(\tau) = \frac{\sum_f I(f \in P_i(\tau) \ \wedge \ f \in T_i)}{\sum_f I(f \in P_i(\tau))} \tag{3.1}$$

$$rc_i(\tau) = \frac{\sum_f I(f \in P_i(\tau) \ \wedge \ f \in T_i)}{\sum_f I(f \in T_i)} \tag{3.2}$$

where $I(f)$ is the standard indicator function, f is a GO term, T_i is the set of true GO terms (true set) for protein i and $P_i(\tau)$ is the set of predicted GO terms for protein i with confidence score greater than or equal to τ. f ranges over the GO hierarchy, excluding the root terms.

The precision-recall space is then generated by averaging precision and recall across all targets at a given threshold. The average precision and recall at a fixed threshold τ were calculated as

$$pr(\tau) = \frac{1}{m(\tau)} . \sum_{i=1}^{m(\tau)} pr_i(\tau) \tag{3.3}$$

$$rc(\tau) = \frac{1}{n} \sum_{i=1}^{n} rc_i(\tau) \tag{3.4}$$

where n is the total number of targets, $m(\tau)$ is the number of targets $\leq n$, on which at least one prediction has been made above threshold τ.

Each prediction model was characterized by a precision-recall curve $(pr(\tau), rc(\tau))_\tau$ (see Figure 3.3). In order to use a single evaluation metric to compare the per-

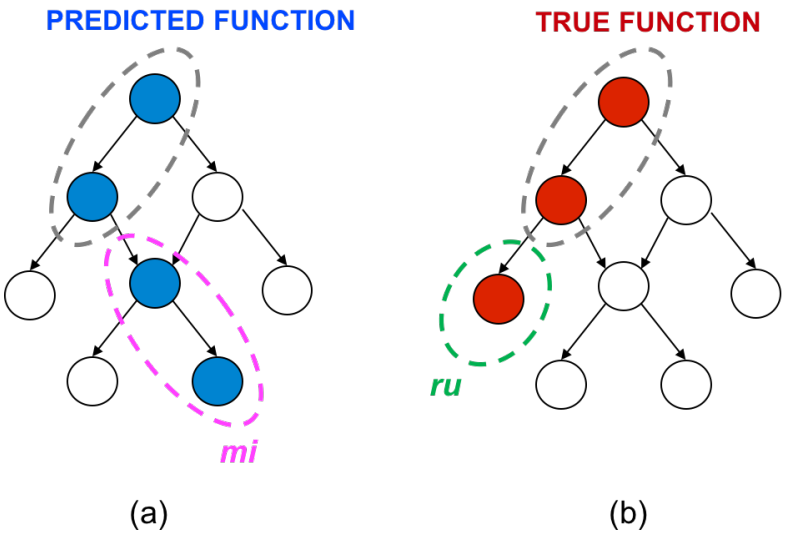


Figure 3.2: Evaluation metrics used for assessment of function prediction methods by CAFA. **(a)** The blue nodes represent the predicted GO terms for particular decision threshold in an ontology. **(b)** The red nodes represent the true GO terms for the corresponding decision threshold in the ontology. The two nodes encircled within grey dashed circles represent the overlap between the predicted and true function sub-graphs. The precision (pr) and recall (rc) for this prediction can be calculated as: $pr = \frac{2}{4} = 0.5$. $rc = \frac{2}{3} = 0.667$. The remaining uncertainty (*ru*) associated with this prediction is the information content of the red node encircled in a green dashed circle while the misinformation (*mi*) is the total information content of the two blue nodes encircled in a magenta dashed circle.

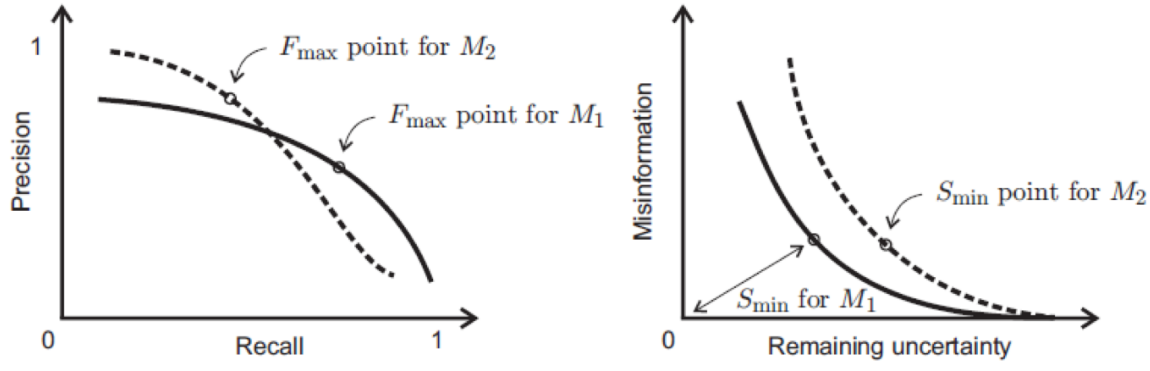


Figure 3.3: This figure shows **(a)** precision-recall or $pr-rc$ curves and **(b)** remaining uncertainty-misinformation $ru-mi$ curves for two function prediction methods M_1 and M_2. The points where F_{max} and S_{min} are achieved are marked as circles in all the curves. Higher F_{max} values and lower S_{min} values can be used to rank function prediction methods. Taken from Friedberg and Radivojac (2016) under CC BY-NC-SA 4.0.

formance of different methods, the maximum F-measure (F_{max}, a harmonic mean between precision and recall, which gives equal emphasis to both) was used over all thresholds. It was calculated as,

$$F_{max} = max\ \tau\{\frac{2.pr(\tau).rc(\tau)}{pr(\tau) + rc(\tau)}\} \tag{3.5}$$

such that a perfect function prediction method would be characterized with F_{max}=1.

(ii) Remaining uncertainty-misinformation ($ru - mi$)

This information-theoretic evaluation metric was introduced by Clark and Radivojac (2013) and was first used in CAFA 2 evaluation to complement the evaluation of function predictions with $pr - rc$ curves due to complexities posed by the structure of biological ontologies and biased or incomplete experimental annotations of biomolecules.

Clark and Radivojac (2013) used a Bayesian network, structured according to the underlying ontology to model the prior probability of a protein's functional annotations and introduced the concepts of misinformation and remaining uncertainty. These terms can be regarded as information-theoretic analogs of precision and recall.

The remaining uncertainty (ru) about a protein's true annotation is regarded as the information about the protein that is not yet provided by the predicted set while misinformation (mi) corresponds to the total information content of the nodes that are incorrect in the predicted set (see Figure 3.2). The information content ($ic(f)$) of a GO term f is estimated in a maximum likelihood manner as the negative binary logarithm of the conditional probability that the GO term f is present in a protein's annotation given that all its parent GO terms are also present.

The average remaining uncertainty(ru) and misinformation (mi) at a fixed decision threshold τ can be calculated as:

$$ru(\tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{f} ic(f).\ I(f \notin P_i(\tau) \wedge f \in T_i) \tag{3.6}$$

$$mi(\tau) = \frac{1}{n}\sum_{i=1}^{n}\sum_{f} ic(f).\ I(f \in P_i(\tau) \wedge f \notin T_i) \tag{3.7}$$

where n is the total number of targets, T_i is the set of true GO terms (true set) for protein i and $P_i(\tau)$ is the set of predicted GO terms for protein i with confidence score greater than or equal to τ.

$ru - mi$ curves are then generated for all targets as the decision threshold is moved from its minimum to its maximum value (see Figure 3.3). A single performance measure, the minimum semantic distance (S_{min}), defined as the minimum distance from the origin to the curve $(ru(\tau), mi(\tau)_\tau$ (Equation 3.8) is then used to rank function prediction methods.

$$S_{min} = min_\tau\{\sqrt{ru(\tau)^2 + mi(\tau)^2} \tag{3.8}$$

where $ic(f)$ is the information content of the GO term f.

3.1.3.2 CAFA 1, 2010-2012

CAFA 1 (Radivojac *et al.*, 2013) organisers provided a set of 48,298 unannotated (lacking any experimental GO terms) proteins from 7 eukaryotic and 11 eukaryotic species as targets to the function prediction community in September 2010. In January 2011, 54 function prediction algorithms associated with 23 research groups had submitted their predictions and the results were released in early 2012. The performance of the top 10 function prediction methods in the Molecular Function Ontology (MFO) is shown in Figure 3.4 where functions predicted by DFX FunFams (listed as Team Orengo) were ranked 7th. The other methods were: Jones-UCL (FunctionSpace) (Cozzetto *et al.*, 2013), Argot2 (Falda *et al.*, 2012), PANNZER (Koskinen *et al.*, 2015), ESG (Chitale *et al.*, 2013), BAR+ (Piovesan *et al.*, 2013), PDCN, Lichtarge Lab (Ward *et al.*, 2009), SIFTER (En-

gelhardt *et al.*, 2006) and dcGO (Fang and Gough, 2013). The performance rankings of the methods often change with benchmark sets, ontologies and evaluation metrics as the performance of a method is dependant on the complex interplay between the prediction method and all the above listed factors (Jiang *et al.*, 2014).

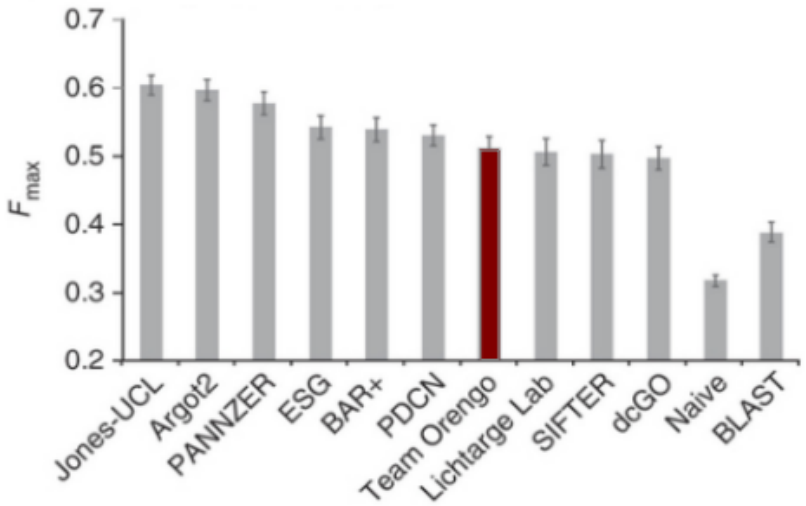


Figure 3.4: The maximum F-measure (F_{max}) for the top 10 performing function prediction methods for Molecular Function Ontology (MFO) where higher values of F_{max} indicates better performance. Adapted from Radivojac *et al.* (2013)

Challenges and limitations of CAFA 1

The first CAFA experiment (Radivojac *et al.*, 2013) was successful in providing an understanding of the trends and performance of existing function prediction methods. At the same time, it also highlighted specific areas of the field which need improvements. However, the most important impact of CAFA 1 was to highlight the major challenges and limitations of automated function prediction to computational biologists, database curators and experimental biologists. Some of these non-trivial challenges included the following:

(i) Protein function is context-based and can be studied from different aspects ranging from biochemical activity to role in pathways, cells, tissues and organisms. However, a function prediction method is often limited by its ability to pro-

cess only certain input data sources (e.g. only eukaryotic proteins) and also by its objective of predicting function in only certain aspects.

(ii) The '**Open World Assumption**' (Thomas *et al.*, 2012; Dessimoz *et al.*, 2013) underlying GO annotations, i.e. function annotations for most proteins are generally incomplete as both experimental annotations and manual curation of annotations are time-consuming and expensive. Consequently, an absence of an annotation does not imply the absence of a function. The failure of the evaluation metrics used in CAFA 1 to account for the Open World Assumption may have led to an overestimation of false-positive predictions in the CAFA evaluation analysis i.e. sometimes correct and highly specific functions predictions may be regarded as false-positives even if proteins have been experimentally annotated only in a more generic manner. This may have significantly affected the results reported in CAFA 1 (Dessimoz *et al.*, 2013).

Additionally, experiments may be biased by the experimenter's choice which can result in the annotations being limited by the scope of experiments. Such experiments are unlikely to determine the entire functional repertoire of proteins and affects our understanding of the protein function space. Function annotations have often been reported to be error-prone, due to experimental interpretations or curator errors (Brenner, 1999; Schnoes *et al.*, 2013). Thus, there may be a number of cases where it may not clear whether a prediction is correct or erroneous. This uncertainty was not captured by the evaluation metrics of CAFA 1 and this led to doubts regarding the reliability of the CAFA 1 results (Dessimoz *et al.*, 2013). In response to this, Jiang *et al.* (2014) studied the effect of incomplete experimental annotations on the reliability of CAFA 1 results by considering function prediction as a structured-output learning problem. They provided theoretical analyses to characterise the impact of missing data on the accuracy of assessments and carried out simulation of the CAFA experiment from which they concluded that although incomplete knowledge can significantly affect assessments, taking available data and realistic assumptions into consideration, the CAFA 1

results are meaningful and reliable Jiang *et al.* (2014).

3.1.3.3 CAFA 2, 2013-2015

CAFA 2 organisers provided 100,816 unannotated or incompletely annotated (experimental GO terms) target protein sequences (from 27 different species - 7 archaeal, 10 bacterial and 10 eukaryotic species) to the protein function prediction community in September 2013. Predictors were asked to predict the function of these proteins by associating gene ontology (GO) terms with the sequences using their methods and upload their results to the CAFA 2 web server before January 2014. The participating methods in CAFA 2 were evaluated on two benchmark sets in two evaluation modes (see Table 3.2) and the results were announced in the Automated Function Prediction Special Interest Group (AFP-SIG) meeting at the Intelligent Systems for Molecular Biology (ISMB)/ European Conference on Computational Biology (ECCB) conference in July 2014.

Table 3.2: Different benchmark sets and evaluation modes used in CAFA 2 to evaluate the performance of the participating methods.

CAFA 2 Benchmark sets	
No-knowledge (NK) benchmark set	Proteins that had no associated experimental GO terms in any ontology in the training database and had accumulated at least one experimental GO term after the accumulation phase. Similar to CAFA1.
Limited-knowledge (LK) benchmark set	Proteins that had experimental GO terms in one or two ontologies in the training database, but not all three, and had accumulated at least one experimental GO term in one or more ontologies after the accumulation phase for which it did not have any experimental terms before.
CAFA 2 Evaluation modes	
Full evaluation (FE) mode	Methods are evaluated on all benchmark proteins and are penalised for not making predictions. Similar to CAFA1.
Partial evaluation (PE) mode	Methods are evaluated on the subset of benchmark proteins for which they have made at least one prediction (as long as they had submitted predictions for at least 5000 targets)

3.2 Aims and Objectives

This chapter discusses the development of an automated function prediction pipeline exploiting the CATH FunFams generated by FunFHMMer, its validation by an in-house CAFA-like benchmark datasets and its independent validation in CAFA 2. It also describes the development of a web server to make function predictions generated by the pipeline available to the scientific community.

The web server was published in:

Das, S., Sillitoe, I., Lee, D., Lees, J. G., Dawson, N. L., Ward, J. and Orengo, C. A. (2015). CATH FunFHMMer web server: protein functional annotations using functional family assignments, *Nucleic Acids Res.*, 43(W1), W148–W153.

The web server was implemented by my colleague and co-author, Dr. Ian Sillitoe.

3.3 Implementation

3.3.1 FunFHMMer pipeline for function annotation

Uncharacterised protein sequences are scanned against a library of HMMs of CATH superfamilies and domain regions are assigned to superfamilies using DomainFinder3 (Yeats *et al.*, 2010) (Figure 3.5(ii)). DomainFinder3 resolves any conflicting or overlapping HMM model matches into a prediction of best non-overlapping matches. The predicted domain sequences are then scanned against the CATH FunFam HMM models for the given superfamily using HMMER3 (Eddy, 2009) and mapped to their best matching FunFam i.e. the model matched with the highest HMM score, provided the inclusion threshold score (described in Section 2.3.5 in Chapter 2) of the respective FunFam model is achieved (Figure 3.5(iii)).

The GO term annotations of that FunFam are then transferred to the query sequence in a probabilistic manner which is calculated as the annotation frequency of a particular GO term amongst the seed sequences of the FunFam (Figure 3.5(iv-v)). The GO term confidence scores are subsequently propagated up the

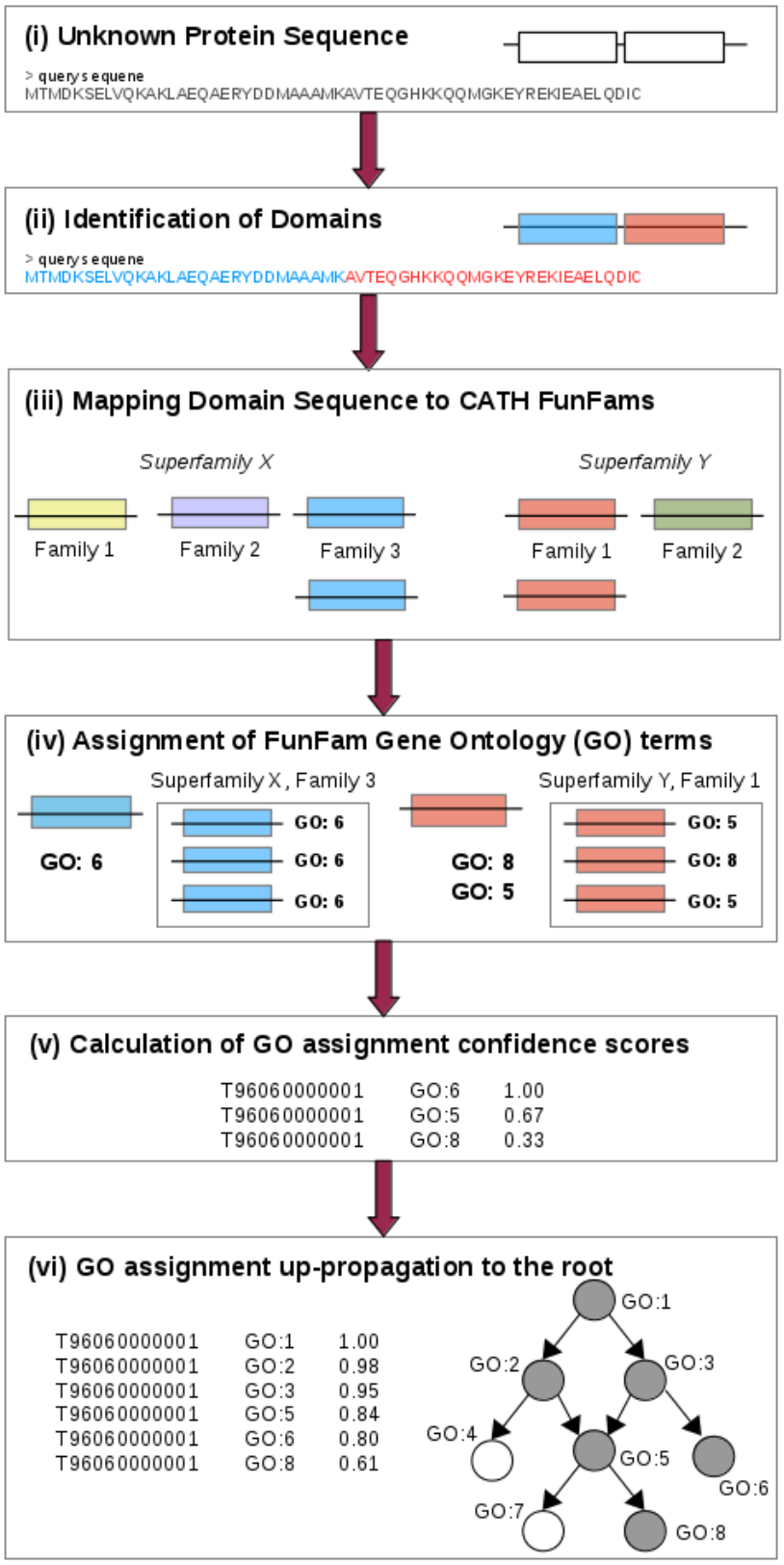


Figure 3.5: Workflow for the FunFHMMer function prediction pipeline. Taken from (Das and Orengo, 2016) under CC BY 4.0.

GO hierarchy (Figure 3.5(vi)). Finally, the non-redundant set of constituent domain GO term assignments for each domain region in the protein sequence, each GO term retaining its highest confidence score, together make up the function predictions for the whole-protein.

The absence of annotations for some query sequences provided by the FunFHMMer function prediction pipeline is most likely due to one of the following reasons: (i) annotations can only be provided for protein families which have one or more known structures classified in CATH (Sillitoe *et al.*, 2015); (ii) query hits are only reported if the sequence match is within the inclusion threshold for the FunFam matched. This is a much stricter criterion than used by many other resources but results in greater precision by preventing misannotations caused by 'over-prediction'. The function predictions by the FunFHMMer pipeline are conservative and focus on higher precision rather than greater coverage.

3.3.2 Benchmarking of function predictions

In order to assess whether the FunFHMMer sub-classification of the domain data in CATH-Gene3D into FunFams improved the functional purity of the FunFams and the ability to use them to transfer functional annotation, we performed a UniProtKB/Swiss-Prot rollback assessment. In this, we compared FunFHMMer against our previous functional classification method, DFX and other domain-family classifications i.e. Pfam and CDD. The domain families in these resources have not been explicitly classified according to function and therefore, the only purpose of including them in the assessment was to determine whether there was any benefit in function annotation transfer from the FunFams.

3.3.2.1 UniProtKB/Swiss-Prot rollback benchmark dataset

A CAFA-style assessment was generated by rolling back the UniProtKB/Swiss-Prot database dated November 2013 to May 2013 (6 months before). The assessment comprised well-annotated sequences which did not have any reported

GO terms (having GO evidence codes: EXP, IDA, IMP, IGI, IEP, TAS or IC) in the Molecular Function Ontology (MFO) in the May 28, 2013 version of UniProtKB/Swiss-Prot, but had MFO annotations associated with them in the November 28, 2013 version. This resulted in a dataset of 1945 proteins. The distribution of leaf MFO term annotations of the assessment proteins is shown in Figure 3.6.

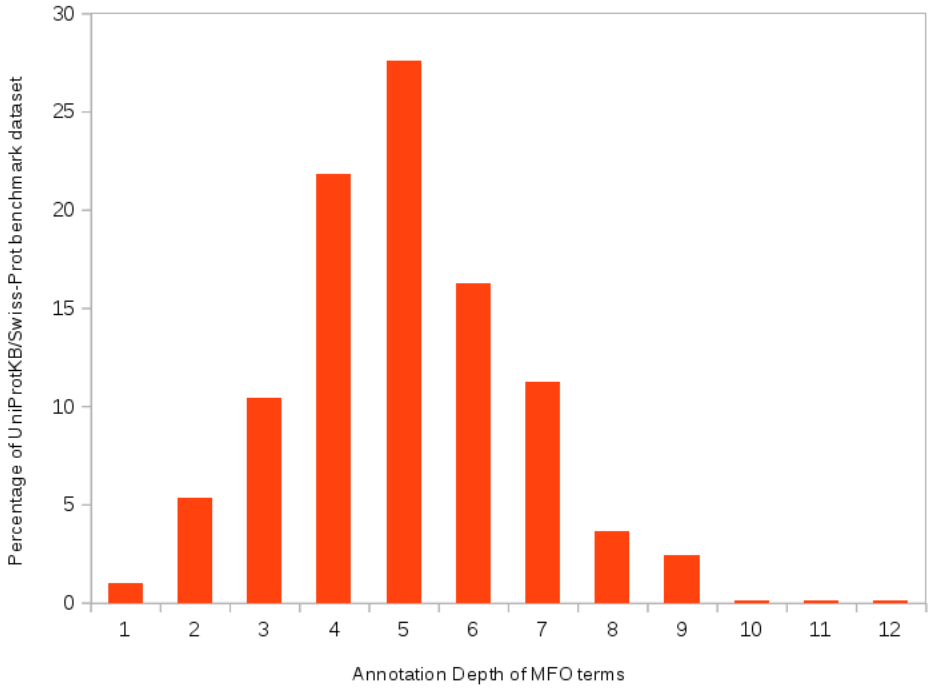


Figure 3.6: Distribution of depths of leaf term annotations of the UniProtKB/Swiss-Prot rollback assessment proteins in Molecular Function Ontology (MFO).

Pfam and CDD were chosen for the assessment as Pfam is the most comprehensive manually curated domain-based resource which is widely used by biologists for functional annotation and CDD is a widely-used comprehensive protein resource that integrates curated protein and protein domain family databases. Each classification protocol was evaluated only on the subset of the assessment dataset for which it predicted at least one GO annotation.

3.3.2.2 Function annotation using Pfam and CDD

Sequence MD5 (a 32 character hexadecimal number) of query sequences was used to map sequences between databases (Smith *et al.*, 2005). The functional annotations assigned by FunFams generated by FunFHMMer were compared to the annotations provided by Pfam (version 27.0) and CDD (version 3.10) family matches.

Pfam families The assessment proteins were scanned against the Pfam (version 27.0) (Finn *et al.*, 2014) family HMM models using HMMER3 (Eddy, 2009). The results were collapsed into a single set of Pfam domain architectures using DomainFinder3 (Yeats *et al.*, 2010) and regions on the proteins are assigned to a Pfam family if the E-value of the match to the HMM is significant (i.e. lower than the inclusion threshold of a Pfam family). The query sequences are assigned the high-quality MFO annotations (extracted from the UniProtKB-GOA annotation file dated May 28, 2013) of annotated sequences in the Pfam family, with a confidence score equal to the annotation frequency of the MFO term amongst all the annotated sequences of that family. This approach is similar to that used for assigning MFO terms and confidence scores to the CATH FunFam matches. The MFO annotations are then propagated up the MFO hierarchy or DAG and the final confidence scores associated with each MFO annotation after up-propagation.

CDD families The assessment proteins were scanned against the CDD (version 3.10) (Marchler-Bauer *et al.*, 2014) family PSSM models using RPS-BLAST. The results were collapsed into a single set of CDD domain architectures using DomainFinder3 (Yeats *et al.*, 2010) and regions on the query proteins are assigned to a CDD family if the E-value of the match is significant (i.e. lower than the domain-specific score thresholds used by the NCBI CD-Search tool to determine whether hits to NCBI-curated domain models are specific or non-specific). The query sequences are assigned high-quality MFO annotations (extracted from the UniProtKB-GOA annotation file dated May 28, 2013) of annotated sequences

in the CDD family with a confidence score equal to the annotation frequency of the MFO term amongst all the annotated sequences of that family. This is similar to the approach used for Pfam family and CATH FunFam matches. The MFO annotations are then propagated up the MFO hierarchy or DAG and the final confidence scores associated with each MFO annotation after up-propagation (described below).

The performance of the functions predicted by different classification methods was measured using Precision-Recall ($pr-rc$) curves by comparing the maximum F-measure (F_{max}) (see Section 3.1.3.1) values.

3.3.2.3 UniProtKB/Swiss-Prot rollback assessment results

The Precision-Recall ($pr - rc$) curves in Figure 3.7 shows the performance of FunFams generated by FunFHMMer in predicting functions for the rollback assessment compared to functions predicted by Pfam families, CDD families and DFX FunFams at different confidence score thresholds ranging from 0-1. Pfam provides predictions for the highest number of sequences (Coverage (C)= 86.5%) in the dataset followed by DFX (C= 75.8%), CDD (C= 74.7%) and FunFHMMer (C= 74%).

From Figure 3.7, we observe that all the methods perform competitively. For predictions with high confidence scores (thresholds $>$ 0.95), Pfam and DFX families i.e broader groupings of protein sequences can predict functions with higher precision than CDD and FunFHMMer. However, for all other predictions with lower confidence scores (thresholds $<$ 0.95), CDD and FunFHMMer perform better with respect to both precision and recall. For this dataset, FunFHMMer gives the highest maximum F-measure (F_{max} = 0.653) than the other family resources (CDD F_{max} = 0.598; DFX F_{max} = 0.595; Pfam F_{max} = 0.581). Figure 3.7 confirms the fact that sub-classifying functionally diverse groups of sequences such as protein domain superfamilies in CATH into functionally coherent sequence groups such as FunFams increases the accuracy of the function predictions for a query se-

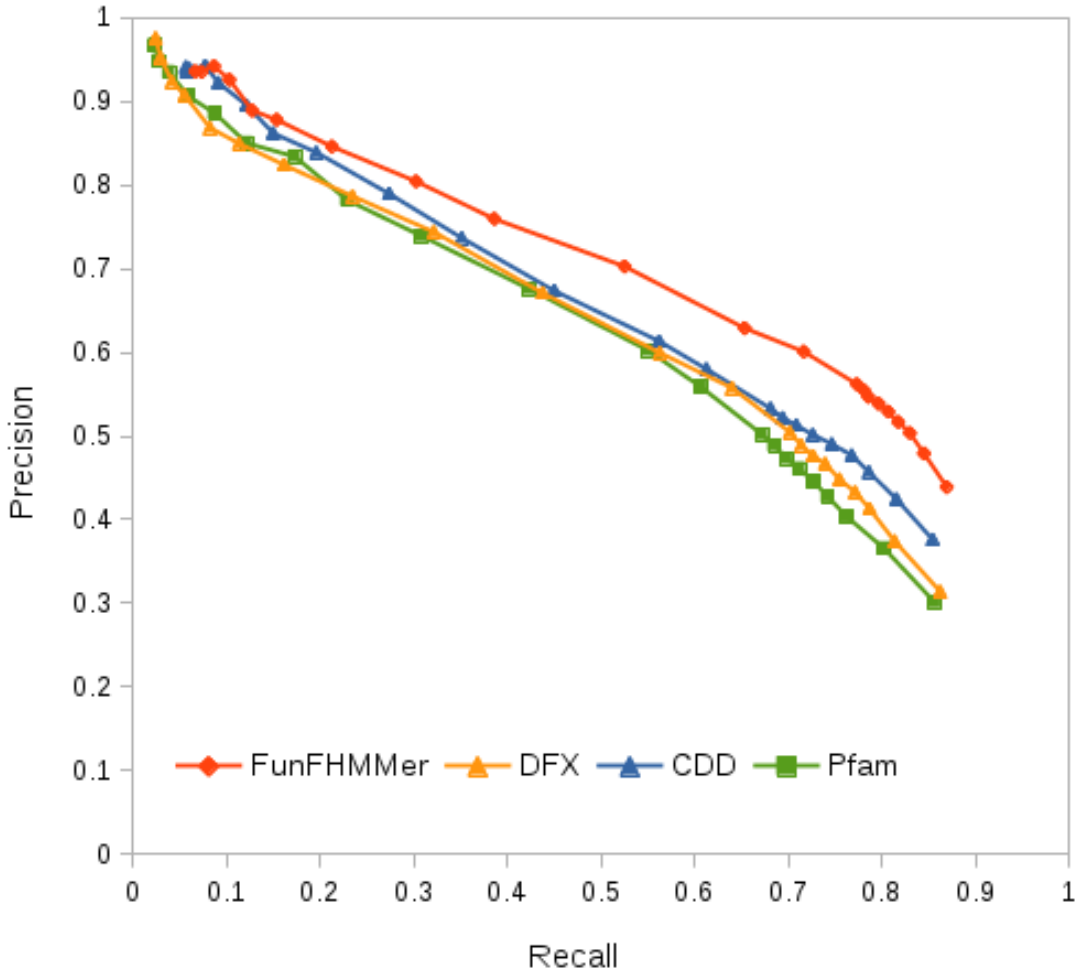


Figure 3.7: Performance of GO annotations predicted by FunFHMMer on the UniProtKB/Swiss-Prot rollback dataset compared to DFX, Pfam and CDD in the Molecular Function Ontology. Taken from (Das and Orengo, 2016) under CC BY 4.0.

quence by restricting the number of homologs from which annotations are transferred to only those which belong to the best matched FunFam. The relative performance of the methods was the same for hard targets of the assessment i.e. those proteins which do not have any functionally annotated relatives with sequence identity $>$ 50% (see Section 3.3.2.4).

FunFHMMer also shows better performance (higher F_{max} value) in predicting protein functions compared to DFX which confirms that, as expected, improved functional sub-classification of CATH superfamilies also improves protein function prediction and that the purity of the FunFams can have a significant impact on

their performance in functional annotation of uncharacterised sequences.

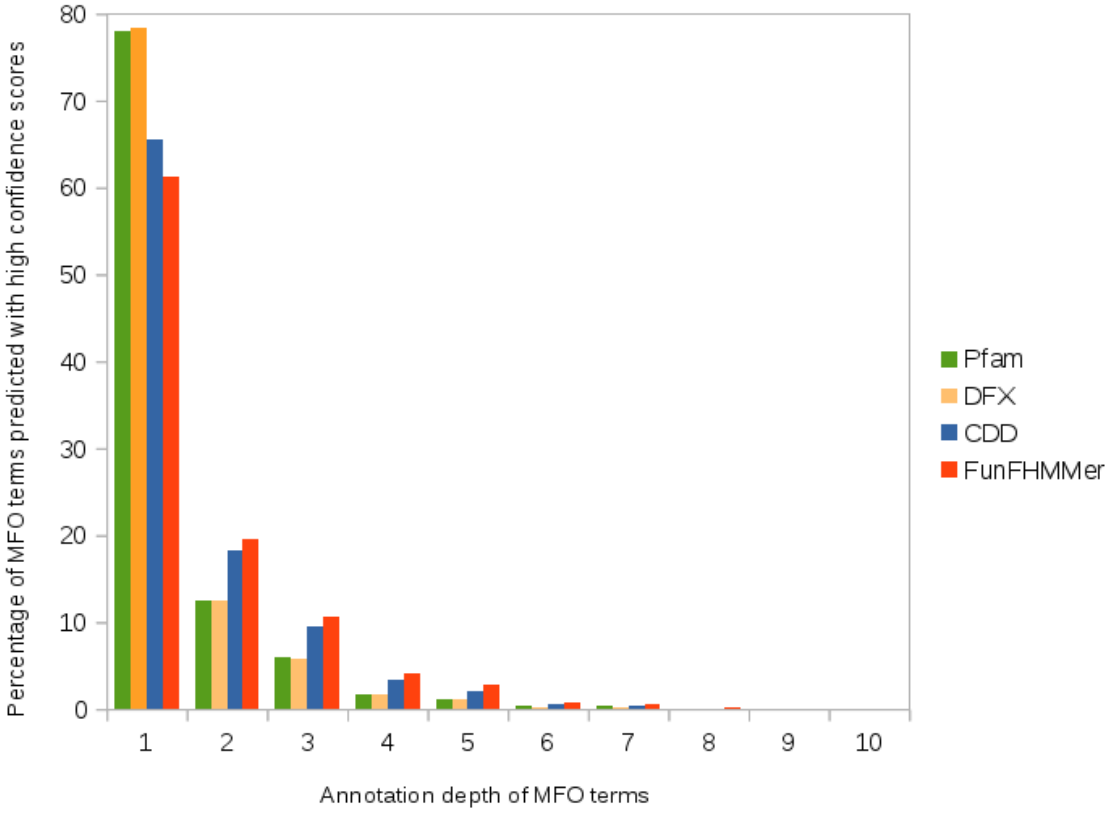


Figure 3.8: Distribution of the high confidence MFO terms of different depths (distance from root of MFO) predicted by the function prediction protocols.

We analysed the predictions with high confidence scores ($>$ 0.95) by all the methods and we observed that indeed Pfam and DFX predict more general MFO terms compared to CDD and FunFHMMer (see Figure 3.8). The depth of the predicted MFO terms (distance to the root of the MFO) was used to indicate whether a MFO term is general or specific the higher the depth, the more specific is the term.

3.3.2.4 Predicting function predictions for hard targets

As many targets in the UniprotKB/Swiss-Prot rollback assessment had very close homologues with functional annotations, which could be easily be recognised by the methods, we also checked the performance of the methods on a subset of the proteins in the dataset which are very hard, i.e. which do not have any functionally

annotated relatives with sequence identity $>$ 50%. This dataset comprised of 553 proteins. The function predictions by the domain-based methods for the hard target benchmark were also compared with GO term predictions obtained from BLAST (version 2.2.29+) (Altschul *et al.*, 1990), the most commonly used tool for function assignment to uncharacterised sequences. The GO annotations from BLAST were predicted by using the annotations of the top annotated BLAST hit for each hard target sequence against the UniProtKB/Swiss-Prot database (dated May 28, 2013) where each MFO annotation is assigned a confidence score equal to 1.

Figure 3.9 shows the performance of FunFHMMer (F_{max} = 0.651, C = 62%), CDD (F_{max} = 0.575, C = 78%), Pfam (F_{max} = 0.555, C = 90%) and DFX (F_{max} = 0.553, C = 72%) on the hard targets of the UniprotKB rollback assessment Dataset. It can be seen that the relative performance of the domain-based prediction methods is similar to the performance on the whole assessment set (see Figure 3.9) and they all outperform BLAST. In Figure 3.9, BLAST is seen to have a limited range of BLAST precision and recall values for different thresholds of confidence scores compared to the family-based function prediction methods. This is because while the family-based methods predict a large number of GO annotations associated with all annotated sequences of the family, BLAST is limited to a single sequence match in this comparison that provides fewer GO annotations for a query sequence. This results in a limited range of confidence scores of the GO annotations predicted by BLAST compared to the family-based prediction methods. Wass *et al.* (2012) had also reported a similar behaviour of BLAST in a Precision-Recall (PR) curve.

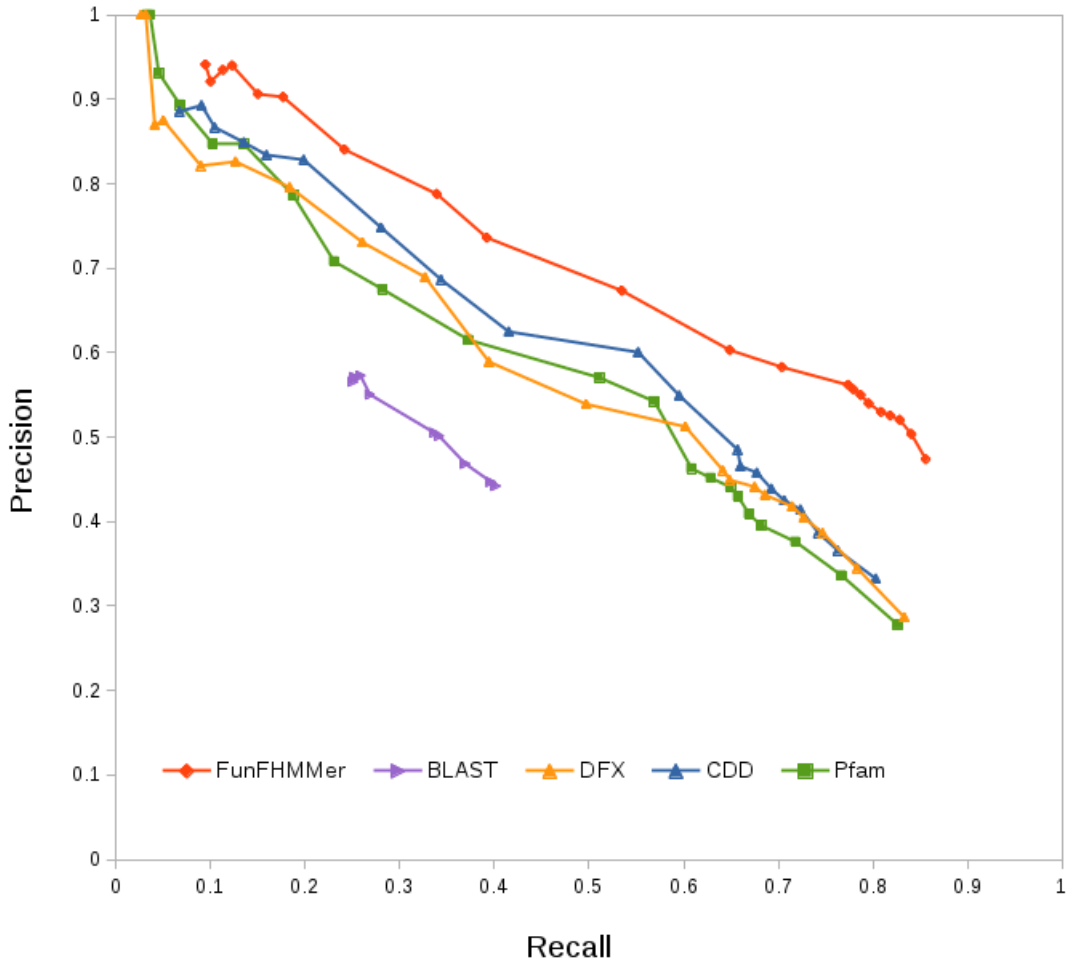


Figure 3.9: Performance of GO annotations predicted by FunFHMMer on hard targets in the UniProtKB/Swiss-Prot rollback dataset compared to DFX, Pfam and CDD in the Molecular Function Ontology.

3.4 FunFHMMer in CAFA 2

3.4.1 Prediction models for CAFA 2

The FunFHMMer function prediction pipeline was used to make function predictions for the 100,816 CAFA 2 targets under the name of the method Orengo-FunFHMMer. Three models utilizing the FunFHMMer function prediction pipeline (see Section 3.3.1) were built. Each of these models are described below:

Orengo-FunFHMMer-1

This model incorporates the FunFHMMer function prediction pipeline in multiple levels using different sets of protein families, such that the subset of targets which do not get assigned any GO terms in the first level are scanned against a different set of protein families in the next level (all levels in Figure 3.10). This was done to improve the coverage of target function predictions. The FunFams generated by FunFHMMer in CATH constitute the first level of the model followed by FunFams generated by FunFHMMer in Pfam, FunFams generated by DFX in CATH, FunFams generated by DFX in Pfam, superfamily level in CATH, family level in Pfam and Naïve predictions. The Naïve predictions were generated by simply predicting all GO terms in an ontology for each target with the relative frequency of the GO term in the training database over all annotated proteins, similar to that used in CAFA 1 (Radivojac *et al.*, 2013). All protein families in this model were annotated with only experimental GO annotations reported in the training database (UniProtKB/SwissProt) which were inherited by query target sequences matching a family.

Orengo-FunFHMMer-2

This model incorporates the FunFHMMer function prediction pipeline in only two levels (the levels shown in green in Figure 3.10) using the protein domain families generated by FunFHMMer in CATH and Pfam, such that the subset of tar-

gets which do not get assigned any GO terms after scanning them against CATH FunFams in the first level are scanned against the Pfam FunFams in the next level. All protein families in this model were annotated with only experimental GO annotations reported in the training database (UniProtKB/Swiss-Prot) which were inherited by query target sequences matching a family. This 'FunFHMMer-FunFam-only' model results in greater precision by preventing misannotations caused by 'over-prediction'.

Orengo-FunFHMMer-3

The structure of this prediction model was exactly same as Orengo-FunFHMMer-1 (all levels in Figure 3.10). However, all protein families in this model were annotated with all GO annotations (i.e. both experimental and automated) reported in the training database (UniProtKB/SwissProt) which were then inherited by query target sequences matching a family.

Orengo-FunFHMMer-MDA

A machine-learning method was built by my colleague, Dr. Jonathan Lees, which used an ExtraTrees Classifier to merge GO term predictions by different protein domain families in a manner similar to the function prediction pipeline of FunFHMMer to make overall GO term predictions for each query sequence (Dr. Jonathan Lees, personal communication). The protein families from which GO term predictions were combined in this method included FunFams generated by FunFHMMer, FunFams generated by DFX, MDA-based CATH-Gene3D FunFams (families consisting of domain sequences sharing the same MDA) and Pfam families. Other information e.g. number of different domains in the protein sequence were also added to the classifier that could effect the confidence of the GO term assignments.

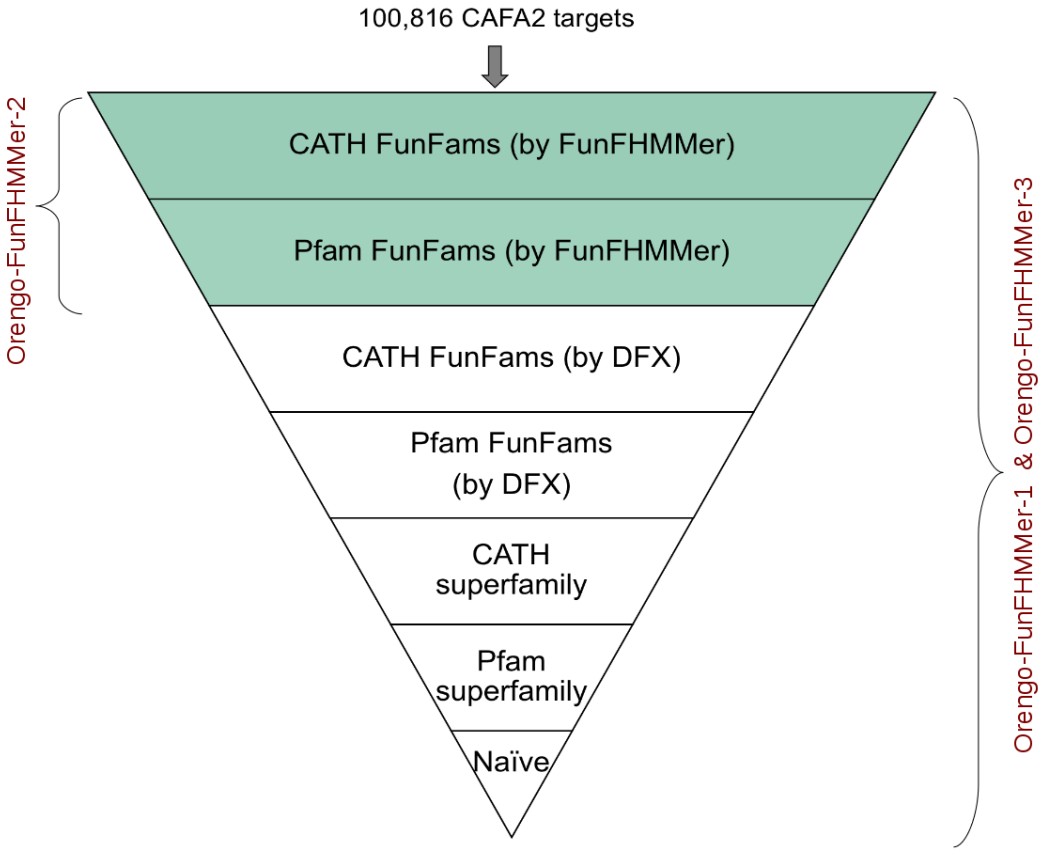
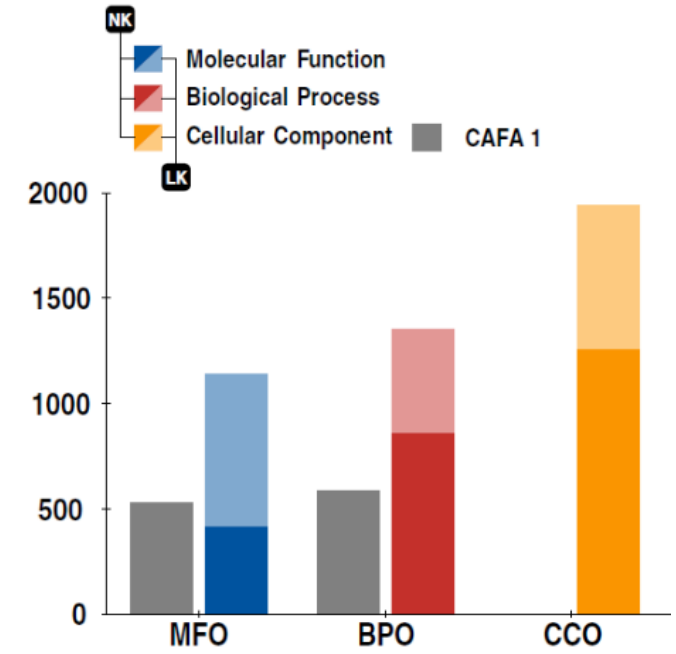


Figure 3.10: The multi-level structure of the FunFHMMer prediction models used for functional annotation of CAFA 2 targets. Orengo-FunFHMMer-1 and Orengo-FunFHMMer-3 used all levels of the prediction model shown here while Orengo-FunFHMMer-2 used only the levels highlighted in green.

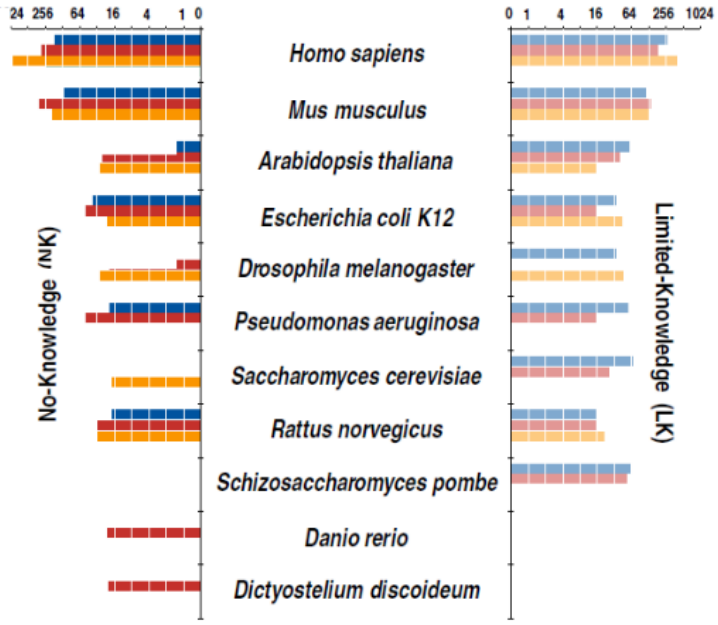
3.4.2 CAFA 2 results

3.4.2.1 General CAFA 2 findings

The CAFA 2 prediction methods were assessed on a subset of the CAFA 2 targets (see Figure 3.11) that accumulated experimental GO terms during the accumulation phase. The assessment of performance of all prediction methods in all evaluation modes is provided by CAFA 2 organizers (Jiang *et al.*, 2016). It is noteworthy that similar to CAFA 1, there was no single best function prediction method for all evaluation modes, benchmarks or ontologies. The performance rankings of the methods often change with ontologies, benchmark sets, evaluation modes and evaluation metrics due to the complex interplay of each of the



(a)



(b)

Figure 3.11: CAFA 2 benchmark. **(a)** The benchmark size for the GO ontologies are shown. **(b)** The number of benchmark sequences for 11 organisms are shown for No-knowledge (NK) and Limited-knolwedge (LK) benchmark types. Adapted from Jiang *et al.* (2016) under CC BY 4.0.

factors involved (see Section 3.1.3). For example, sometimes the differences in the evaluation metric between the top performing methods can be very small in a particular category which may be attributed to insufficiency of currently available data to make predictions for certain targets. However, an analysis of all function prediction methods in different categories using different benchmarks and evaluation modes provides an in-depth analysis of the strengths and weaknesses of each of the methods along with the progress of the methodologies used by the function prediction community to provide accurate and meaningful functional annotations for guiding biological research.

In general, the top methods performed better in predicting MFO terms compared to BPO terms (Jiang *et al.*, 2016). For CCO, the methods did not show any marked improvement over the Naïve method which is most likely due to the frequent use of very general CCO terms such as cytoplasm or nucleus in annotations. The performance of the top methods on both easy and hard targets for MFO and BPO was found to be similar unlike BLAST, whose performance was severely affected when assessed on hard targets. Moreover, the top methods in CAFA 2 were found to outperform the top methods in CAFA 1 based on their assessment on an overlapping target set. This increase in performance can be regarded as a combined effect of the increasing experimental annotation data and improvement of the prediction methods (Jiang *et al.*, 2016).

3.4.2.2 Top ranking function prediction methods in CAFA 2

The performance of the function prediction methods in CAFA 2 was assessed by the CAFA 2 organisers. All the figures in this Section have been taken from the figures provided by the CAFA 2 organisers (Iddo Friedberg, personal communication) to the Automated Function Prediction community that specify the name of the model for the top ranking methods. These figures are also available in the Supplementary Material of Jiang *et al.* (2016), however, the specific models of each of the top ranking methods are not mentioned in these figures.

Molecular Function Ontology (MFO)

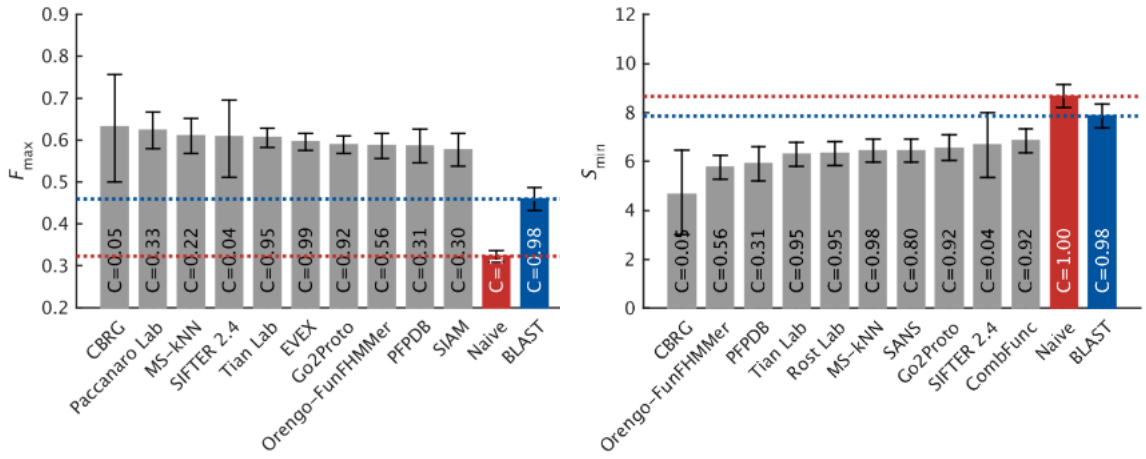


Figure 3.12: Top 10 CAFA 2 methods for the analysis of **all** targets in the **NK** benchmark set using **partial** evaluation mode in the **Molecular Function Ontology** (MFO) evaluated using **(a)** F_{max} **(b)** S_{min}

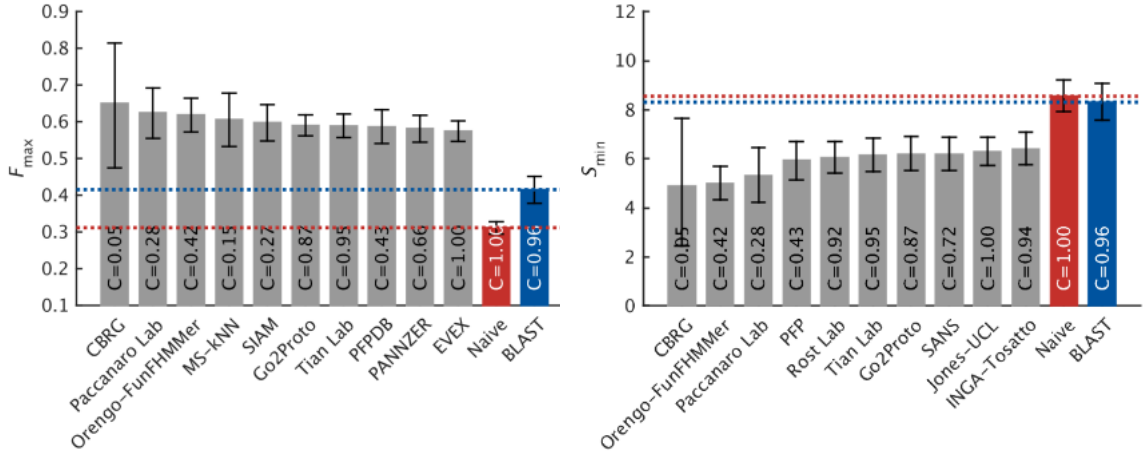


Figure 3.13: Top 10 CAFA 2 methods for the analysis of **hard** targets in the **NK** benchmark set using **partial** evaluation mode in the MFO evaluated using **(a)** F_{max} **(b)** S_{min}

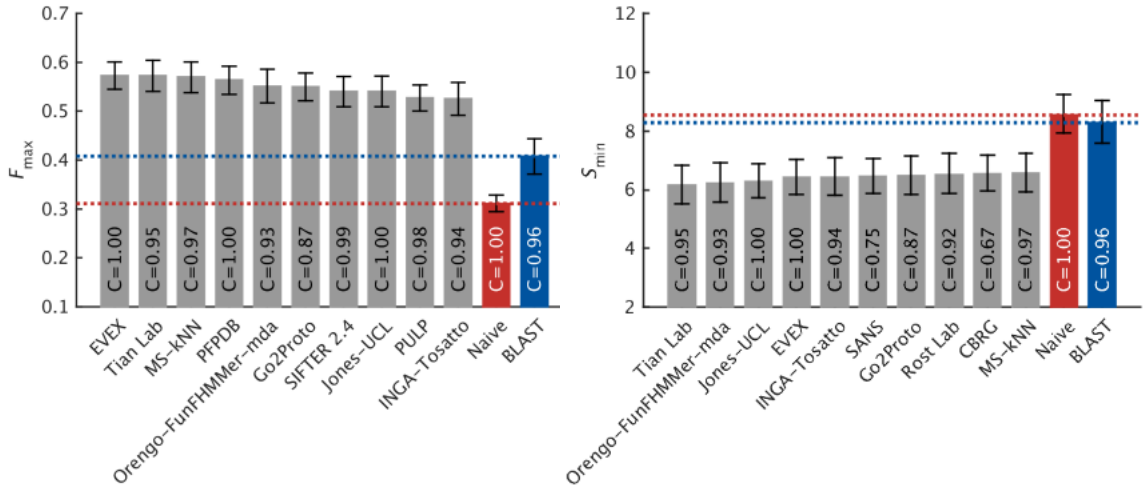


Figure 3.14: Top 10 CAFA 2 methods for the analysis of **hard** targets in the **NK** benchmark set using **full** evaluation mode in the MFO evaluated using **(a)** F_{max} **(b)** S_{min}

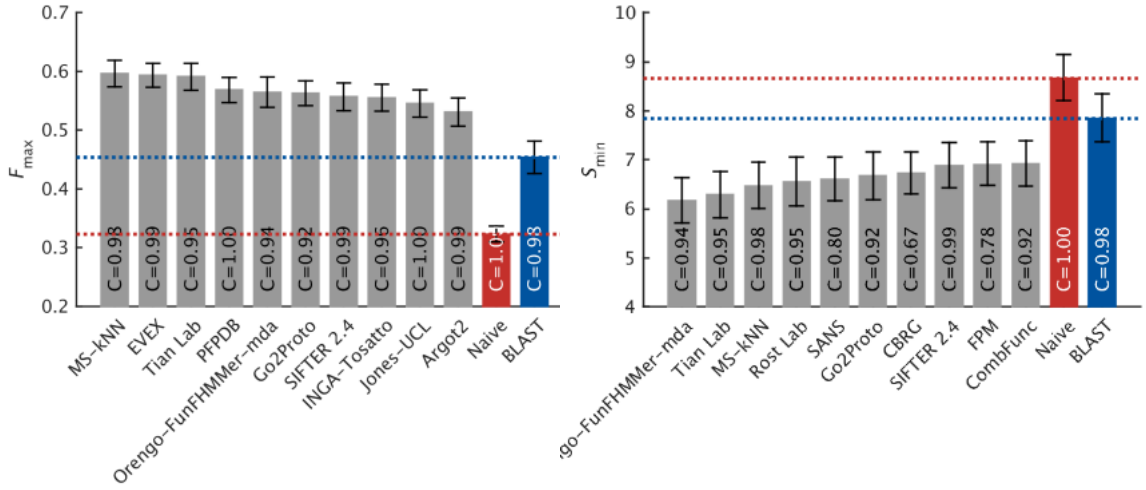


Figure 3.15: Top 10 CAFA 2 methods for the analysis of **all** targets in the **NK** benchmark set using **full** evaluation mode in the MFO evaluated using **(a)** F_{max} **(b)** S_{min}

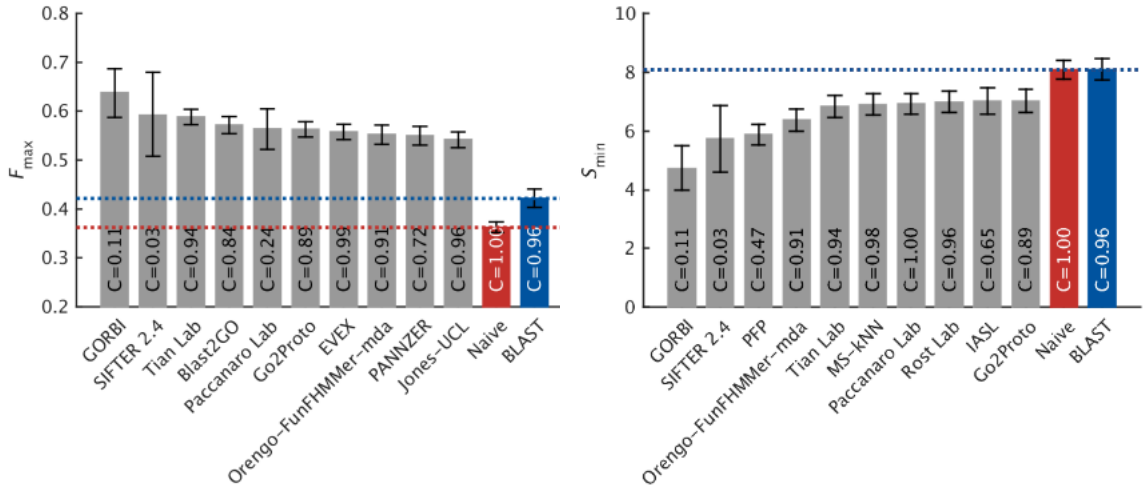


Figure 3.16: Top 10 CAFA 2 methods for the analysis of **all** targets in the **LK** benchmark set using **partial** evaluation mode in the MFO evaluated using **(a)** F_{max} **(b)** S_{min}

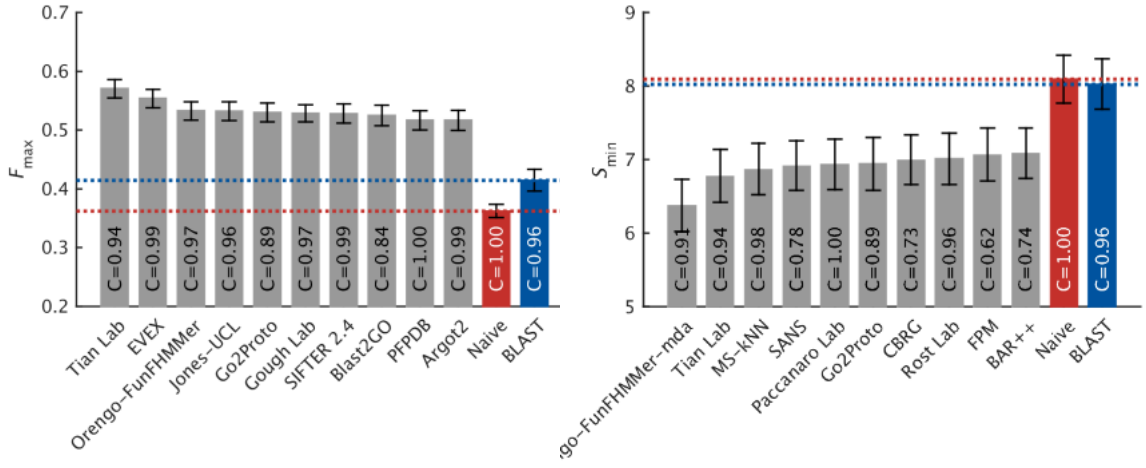


Figure 3.17: Top 10 CAFA 2 methods for the analysis of **all** targets in the **LK** benchmark set using **full** evaluation mode in the MFO evaluated using **(a)** F_{max} **(b)** S_{min}

Biological Process Ontology (BPO)

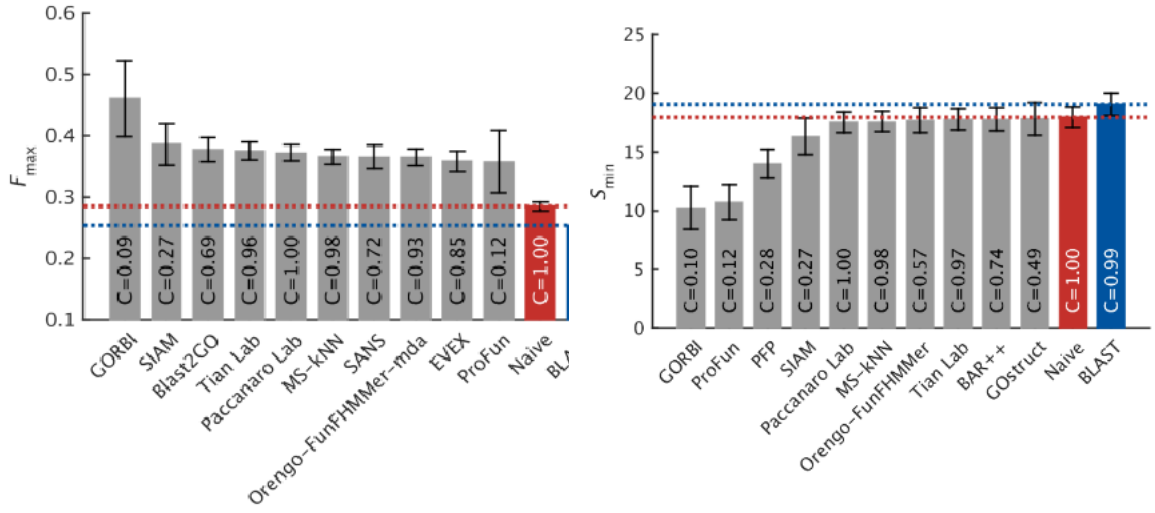


Figure 3.18: Top 10 CAFA 2 methods for the analysis of **all** targets in the **NK** benchmark set using **partial** evaluation mode in the **Biological Process Ontology** (BPO) evaluated using **(a)** F_{max} **(b)** S_{min}

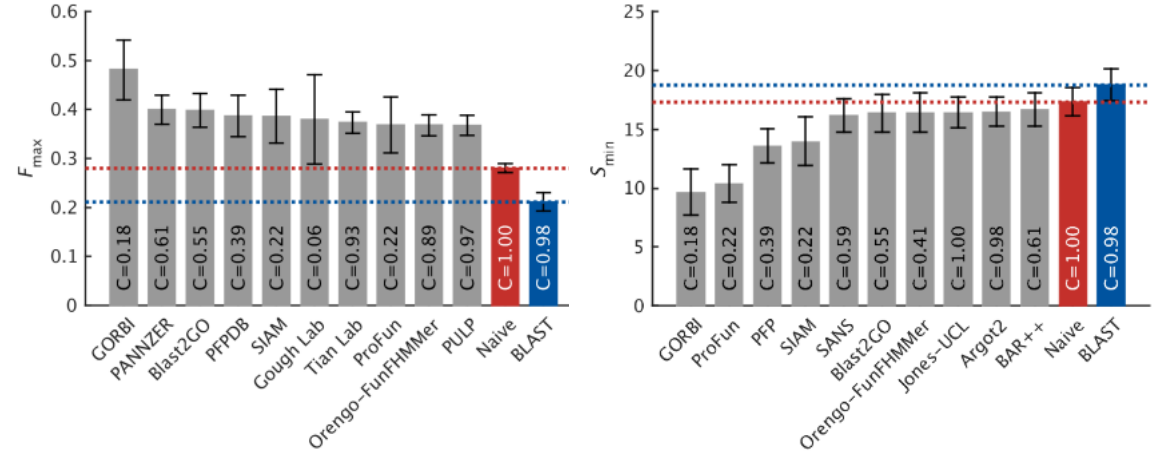


Figure 3.19: Top 10 CAFA 2 methods for the analysis of **hard** targets in the **NK** benchmark set using **partial** evaluation mode in the **BPO** evaluated using **(a)** F_{max} **(b)** S_{min}

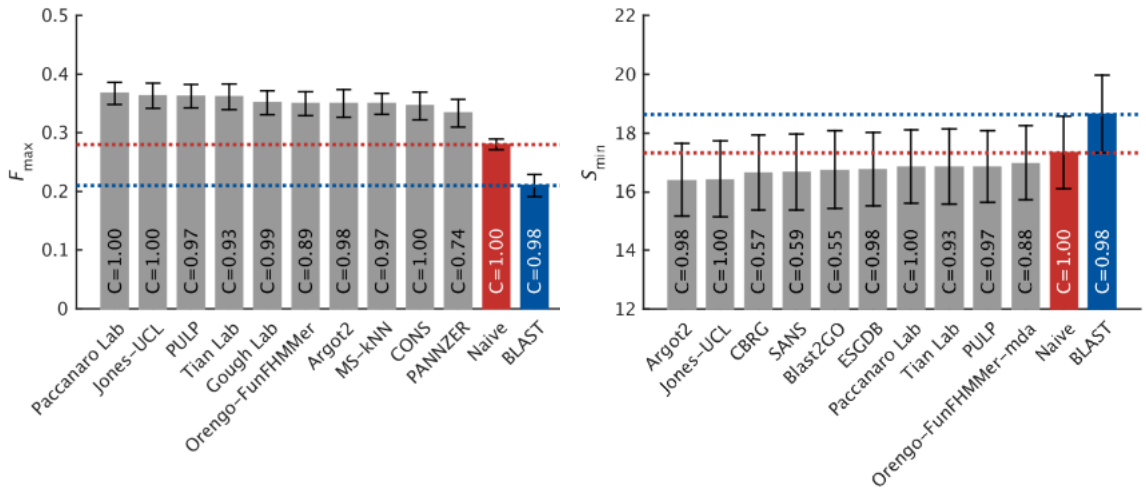


Figure 3.20: Top 10 CAFA 2 methods for the analysis of **hard** targets in the **NK** benchmark set using **full** evaluation mode in the BPO evaluated using **(a)** F_{max} **(b)** S_{min}

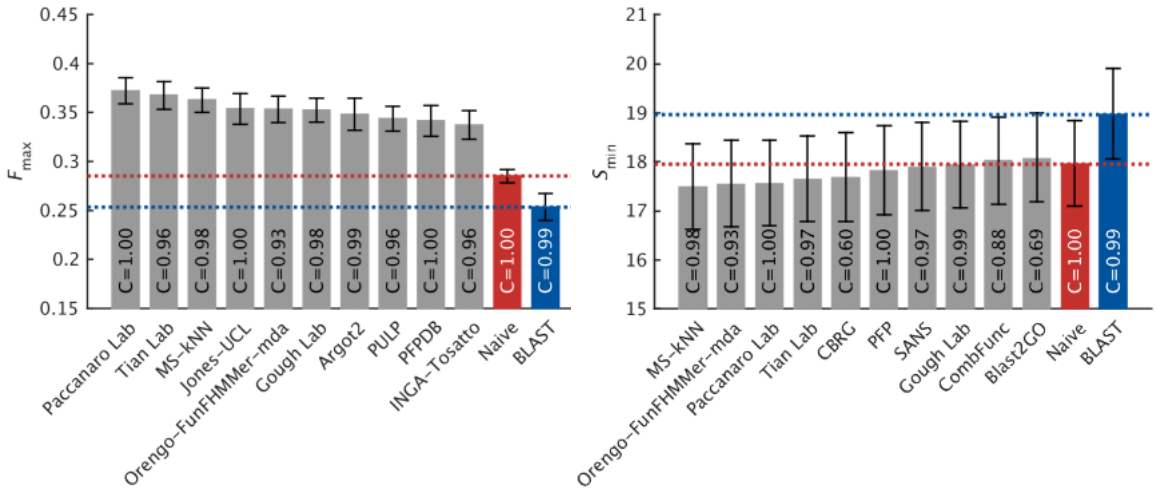


Figure 3.21: Top 10 CAFA 2 methods for the analysis of **all** targets in the **NK** benchmark set using **full** evaluation mode in the BPO evaluated using **(a)** F_{max} **(b)** S_{min}

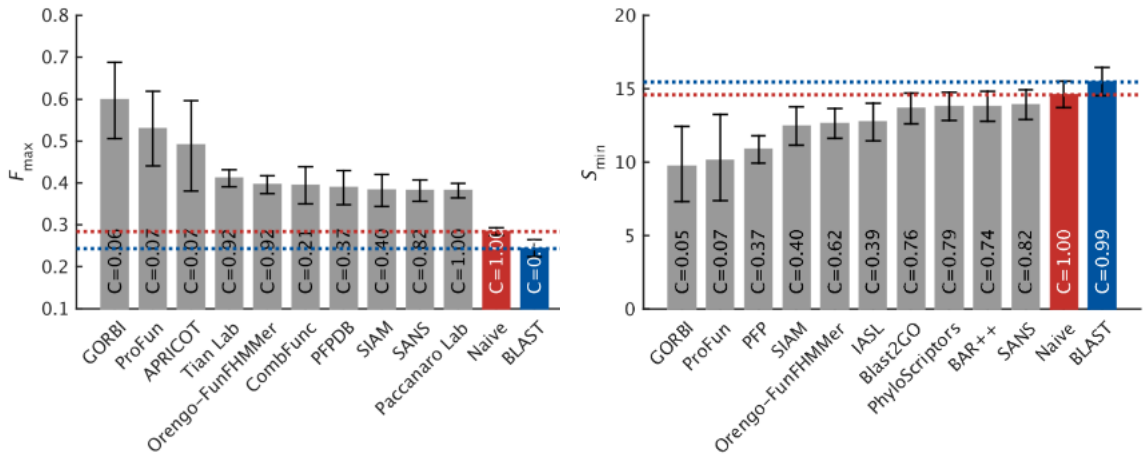


Figure 3.22: Top 10 CAFA 2 methods for the analysis of **all** targets in the **LK** benchmark set using **partial** evaluation mode in the BPO evaluated using **(a)** F_{max} **(b)** S_{min}

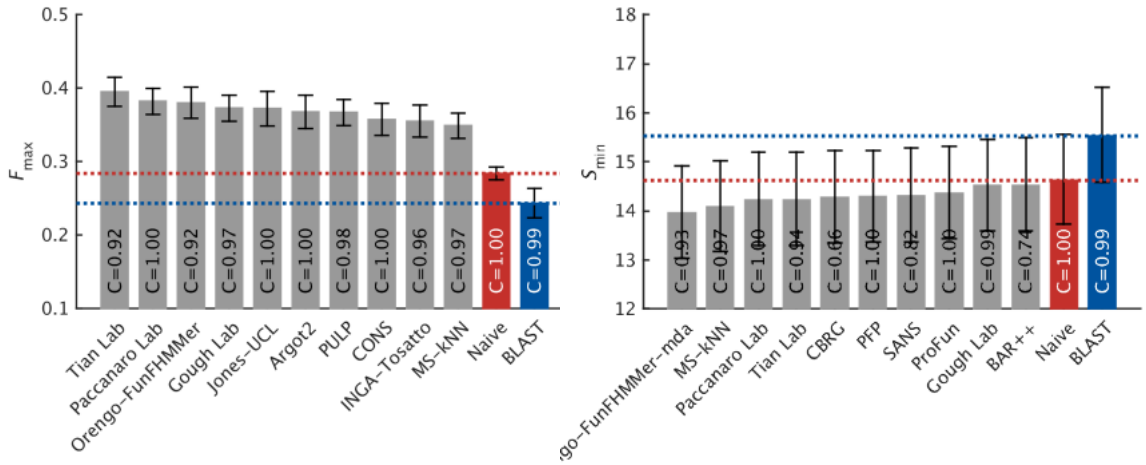


Figure 3.23: Top 10 CAFA 2 methods for the analysis of **all** targets in the **LK** benchmark set using **full** evaluation mode in the BPO evaluated using **(a)** F_{max} **(b)** S_{min}

3.4.2.3 Performance of FunFHMMer-based methods in CAFA 2

Orengo-FunFHMMer-1 and Orengo-FunFHMMer-3 generated at least 1 GO term prediction for all 100,816 CAFA 2 targets since Naïve prediction formed the last level for both the models for any target sequence unannotated by previous levels of the models. However, the CATH FunFams are used to predict functions for $>$71.5% of the targets using these two models. On the other hand, Orengo-FunFHMMer-2 generated predictions for 59,114 (58.6%) CAFA 2 targets.

Figures 3.12 and 3.18 shows the performance of the FunFHMMer-based methods compared to other top function prediction methods for all CAFA 2 targets in the No-knowledge (NK) benchmark set using the partial evaluation mode in the Molecular Function Ontology and Biological Process Ontology respectively. The performance of the FunFHMMer-based prediction models for all other benchmark sets using different evaluation modes and metrics that have been provide by the CAFA 2 organisers (Jiang *et al.*, 2016) are shown in Section 3.4.2.2. In all CAFA 2 performance figures, the organisers presented only the best performing method submitted under one principal investigator. As a result, either the Orengo-FunFHMMer method or the Orengo-FunFHMMer-MDA method appears in each figure, depending on which method performed better using a particular evaluation metric.

Overall, it can be seen that the Orengo-FunFHMMer model generally performs better when it is analysed on the partial subset of CAFA 2 targets i.e when it is assessed only on those targets for which it made at least one function prediction and no penalisations are made for targets which did not receive any predictions. On the other hand, Orengo-FunFHMMer-MDA performs better in full CAFA 2 target set as it provides predictions for a larger number of targets. Moreover, out of the three models of Orengo-FunFHMMer, the Orengo-FunFHMMer-2 model was found to perform better for partial targets benchmark sets, which is likely due to its more specific function predictions, while Orengo-FunFHMMer-3 model performed better for all targets benchmark sets because of its better coverage. This can be

easily explained by the fact that the Orengo-FunFHMMer pipeline is more conservative (Orengo-FunFHMMer-2 being its most conservative model) and hence, more precise, compared to the Orengo-FunFHMMer-MDA pipeline which uses machine learning to combine the individual function predictions from CATH FunFams and other protein families including DFX FunFams, MDA-based FunFams and Pfam families to generate overall predictions, providing greater coverage. Also, it can be seen that all FunFHMMer-based methods rank higher when evaluated using S_{min} compared to F_{max} for the same benchmark set, which supports the fact that FunFHMMer-based methods provide more specific (i.e. more informative) GO terms compared to other top function prediction methods in CAFA 2.

3.5 The FunFHMMer Web Server

The FunFHMMer web server is available at `http://www.cathdb.info/search/by_funfhmmer`. It is implemented as a Perl-based web application that interacts with a custom distributed queueing system (using beanstalkd as a simple and fast work queue and memcached to provide a distributed storage model for the results). Individual scans should only take seconds and the queueing system enables this performance to scale well.

3.5.1 Input

The FunFHMMer web server can be queried using a protein sequence in the FASTA format or by entering UniprotKB or GenBank sequence identifiers as input in the text area on the web page (Figure 3.24). The length of the input sequence is not limited and the search for function predictions for query sequences by FunFHMMer is typically very fast ($<$ 1 minute). However, it may take up to several minutes for very long sequences. A fully documented application programming interface (API) is also provided for interfacing the FunFHMMer search from within any software application.

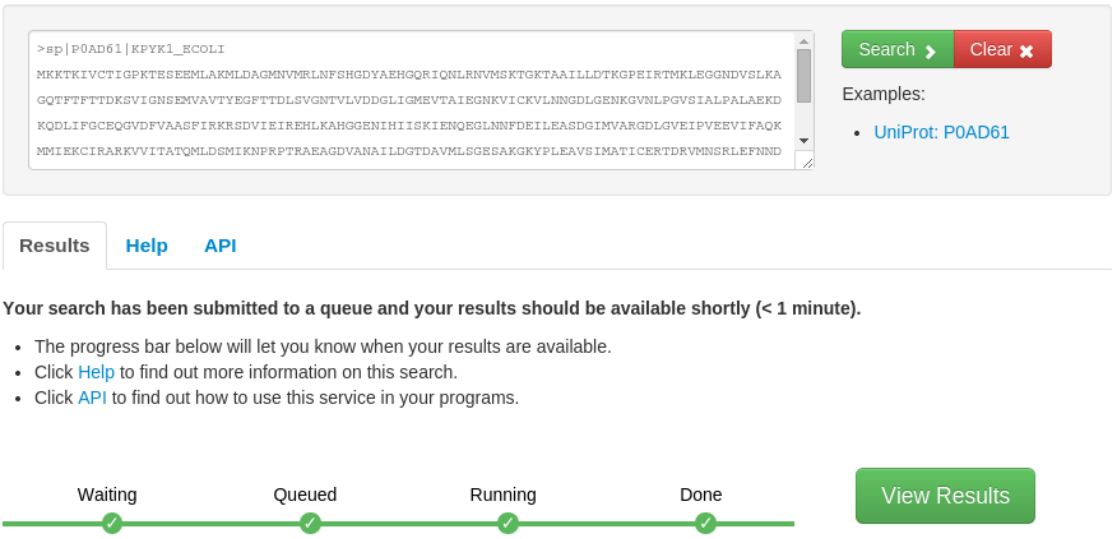
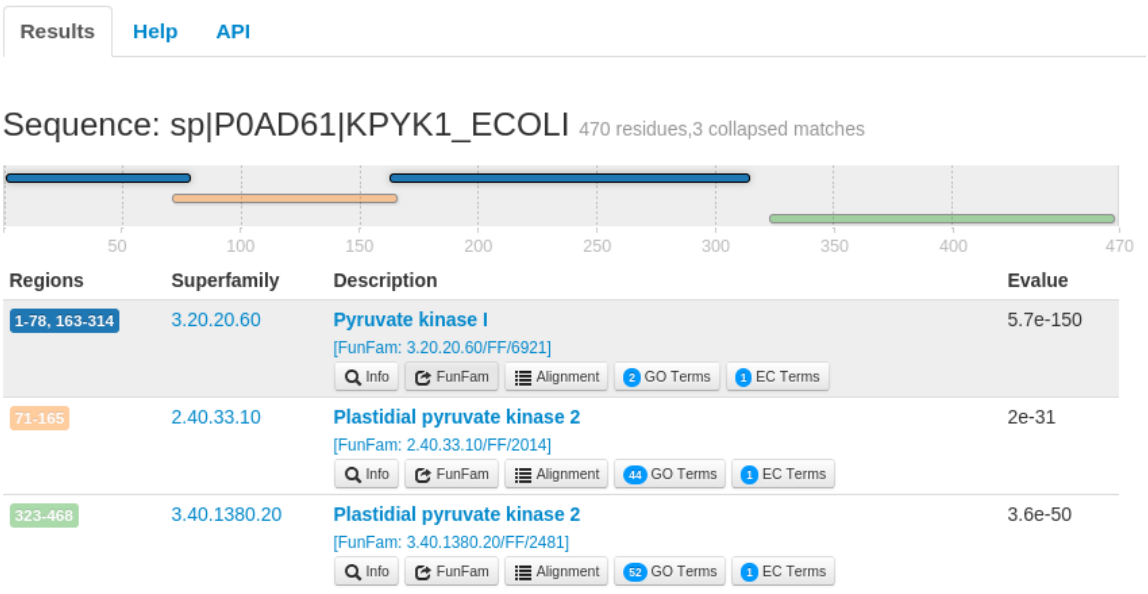


Figure 3.24: The FunFHMMer web server query page.

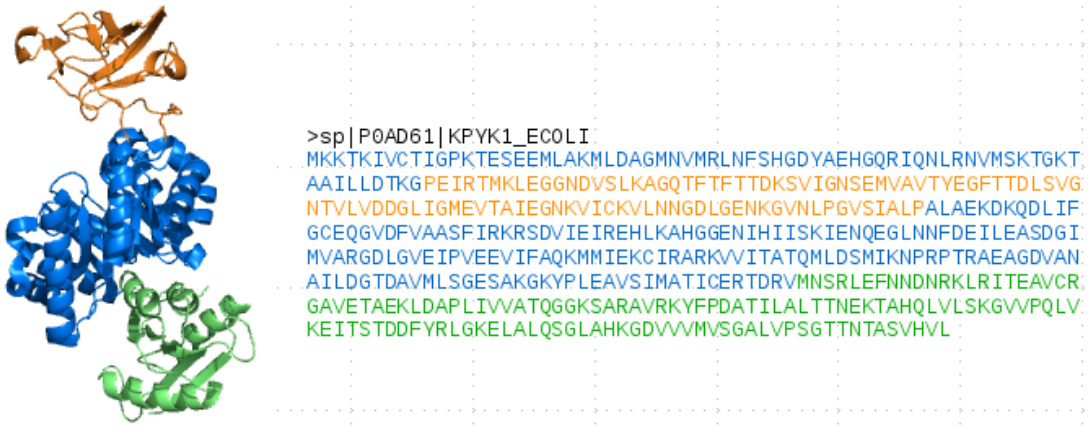
3.5.2 Output

The output of the web server provides the multi-domain architecture of the query sequence along with CATH domain superfamily and FunFam assignments for each domain identified within the query sequence (Figure 3.25a). The EC and GO annotations for each of the predicted FunFams are displayed in tables along with their annotation frequency. The GO annotation table can be visualised using the REViGO (Supek *et al.*, 2011) web server using the link provided in the results page.

For example, for the UniprotKB sequence P0AD61, FunFHMMer assigns three structural domains (see Figure 3.25) along with their significant E-values (i.e. E-values $<$ FunFam inclusion threshold). The first domain is discontinuous (shown in blue) and matches FunFam 6921 in CATH superfamily 3.20.20.60, the second continuous domain (shown in yellow) matches FunFam 2014 in the CATH superfamily 2.40.33.10 and the third continuous domain matches FunFam 2481 in the CATH superfamily 3.40.1380.20. The description of the FunFams is automatically generated by text-mining the UniprotKB descriptions of the sequence relatives for each FunFam. These terms may reflect the function of the whole protein rather than the function of the individual domain. For each CATH FunFam



(a)



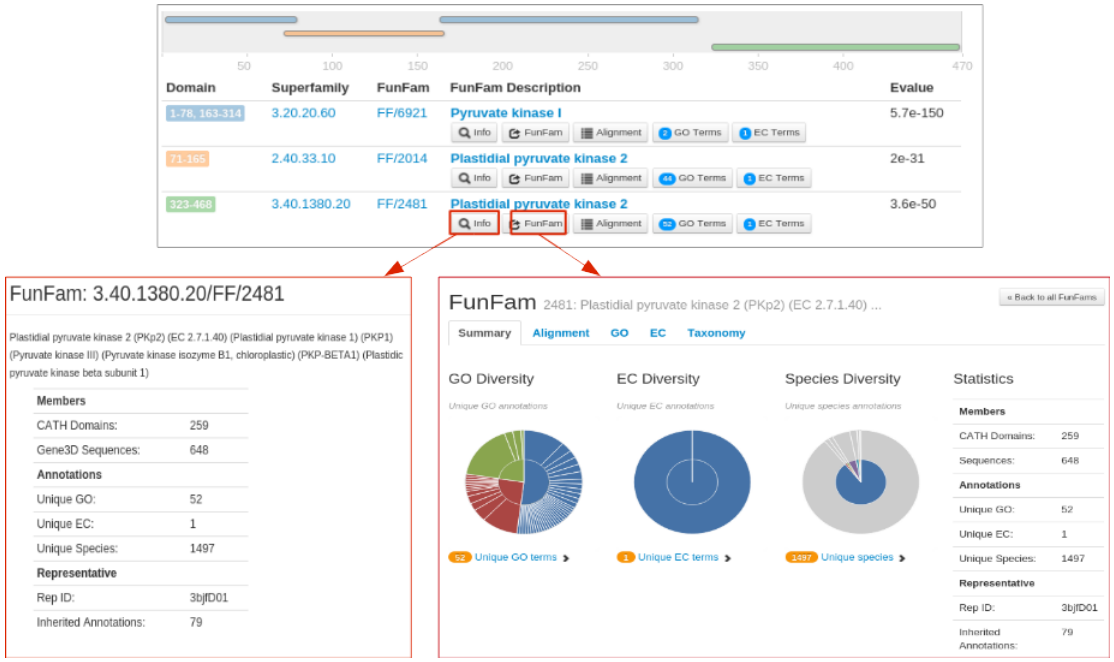
(b)

Figure 3.25: (a) Example of FunFHMMer web server results. CATH superfamilies and FunFams have been identified within the example UniprotKB sequence P0AD61 submitted to the FunFHMMer web server. Functional information can be retrieved through GO Terms and EC Terms buttons. **(b)** Structural domains identified by CATH shown in different colours in the structure (PDB: 1E0T) associated with P0AD61. The blue domain is discontinuous whereas the yellow and green domains are continuous domains.

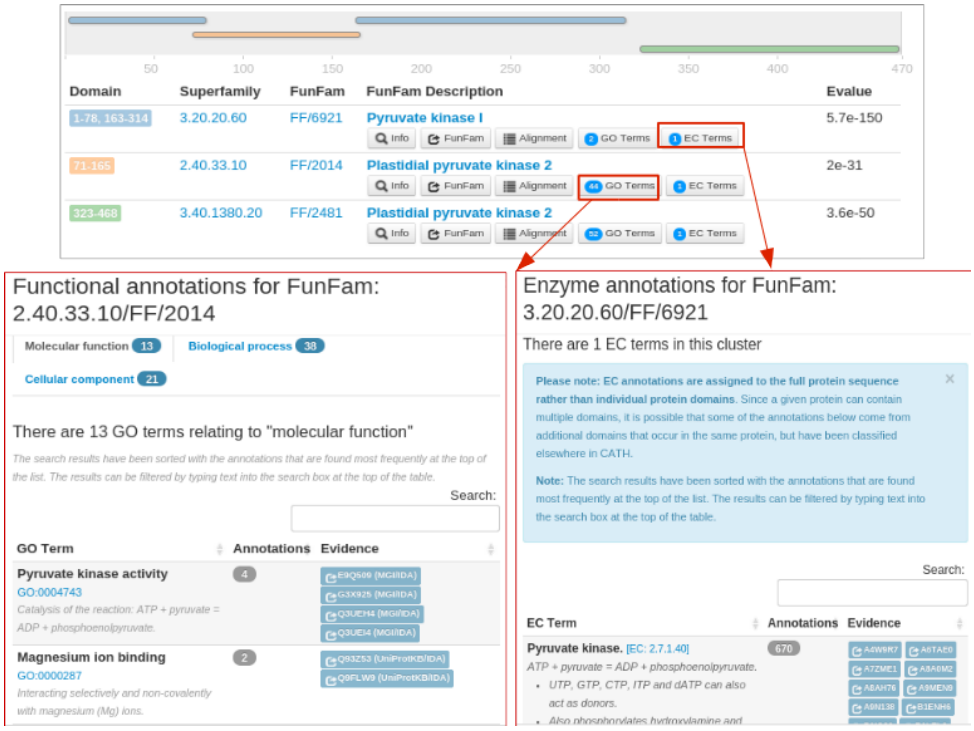
match, the 'Info' button provides a brief description about the FunFam. To know more about a FunFam, the 'FunFam' button directs the user to the CATH FunFam web page which can provide useful functional and structural information (see Figure 3.26a). For example, information on highly conserved positions, highlighted in green, in the FunFam multiple-sequence alignment identified using Scorecons Valdar (2002) is shown on a representative protein domain structure to highlight residues that are expected to be functionally important.

The GO annotations and the EC annotations corresponding to each domain are available via the 'GO Terms' and 'EC Terms' buttons, along with their annotation frequencies (see Figure 3.26b). The web server also provides a detailed help page to assist users in understanding how to submit query sequences to the FunFHMMer web server and interpret its output. The Alignment button for each FunFam shows the alignment of the query sequence domain region aligned to the CATH FunFam HMM match using HMMER3 (Eddy, 2009). For example, the Figure 3.27 shows the alignment of the third predicted structural domain in the query sequence (residues 323-468) to FunFam 2481 in the CATH superfamily 3.40.1380.20.

A query protein sequence can have multiple hits to different but related FunFams within a single CATH superfamily. For example, the yeast Pyruvate decarboxylase (UniprotKB Accession number: P06169) is a TPP-dependant enzyme which consists of three domains: a pyrimidine (PYR) binding domain, a transhydrogenase dIII - (TH3) domain and a pyrophosphate (PP) binding domain, where the PP and the PYR domains are known to be evolutionarily related (Costelloe *et al.*, 2008). The FunFHMMer web server results for UniprotKB sequence P06169 matches two hits in the CATH superfamily 3.40.50.970 (with different FunFam matches) and one hit to the CATH superfamily 3.40.50.1220 (Figure 3.28). The two hits in 3.40.50.970 is because the PP and PYR domains are homologous domains which result from a gene duplication during evolution have been classified into the same CATH superfamily - a relationship confirmed by structural data.



(a)



(b)

Figure 3.26: Example of FunFHMMer web server result pages for UniProtKB sequence P0AD61 showing details of **(a)** the FunFams assigned to it and **(b)** EC and GO annotations predicted by FunFHMMer.

Figure 3.27: The FunFHMMer web server displays alignments of predicted structural domains with the FunFam model assigned to it that are generated by HMMER3 (Eddy, 2009). This figure shows the alignment of the third predicted structural domain in the example UniProtKB sequence P0AD61 (residues 323-468) to the FunFam assigned to it.

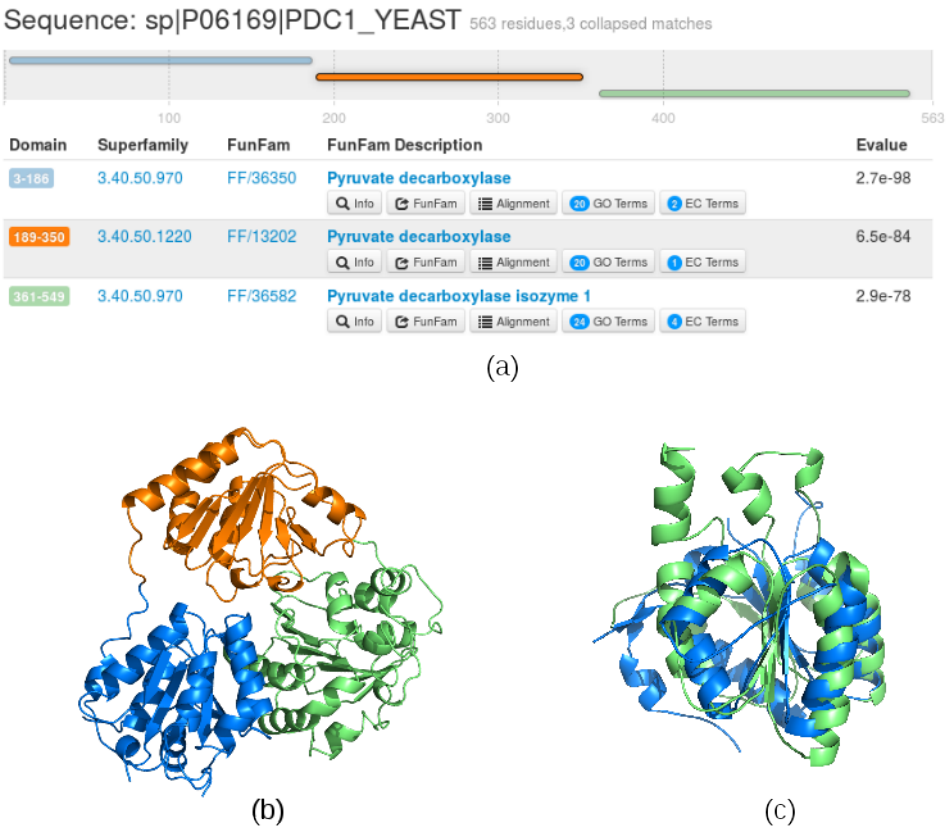


Figure 3.28: **(a)** FunFHMMer web server results for the UniprotKB protein sequence P06169 (yeast Pyruvate decarboxylase) are shown. The structural domains identified by CATH are shown - first domain assigned to superfamily 3.40.50.970 and FunFam 36350 (PYR domain) is coloured blue, second domain assigned to superfamily 3.40.50.1220 and FunFam 13202 (TH3 domain) is coloured orange and third domain assigned to superfamily 3.40.50.970 and FunFam 36582 (PP domain) is coloured green. **(b)** The structure of the yeast Pyruvate decarboxylase (PDB: 1PVD) is shown where the structural domains are coloured as in subfigure (a). **(c)** The structural alignment (performed by SSAP) of the PYR and PP domains of the yeast Pyruvate decarboxylase illustrates that the two domains are homologous.

3.6 Conclusions and Discussion

The function of proteins in a cell is very complex and is tightly regulated by various mechanisms. Use of sequence and structure homology to experimentally characterised proteins provides important clues regarding the function of a protein. However, in many cases, knowledge of gene or amino acid sequence and/or three-dimensional structure of a protein is not always sufficient to infer function of a protein. This is one huge challenge for protein function prediction that a large number of different protein functions is sometimes encoded by a single gene or by the same amino acid sequence. For example, (i) a gene can sometimes encode multiple proteins through alternative splicing, the function of a protein can be affected by (ii) post-transcriptional (RNA editing) and (iii) post-translational modifications, (iv) a gene can sometimes lose its ability to encode a protein by accumulation of multiple mutations (pseudogene) (v) proteins can evolve new functions (moonlighting proteins) and (vi) sometimes the function of a protein is limited to a specific site or location within a cell. Furthermore, protein function is context-based and can be studied from different aspects ranging from biochemical activity to the role of the protein in pathways, cells, tissues and organisms.

Apart from the complexity of inference of protein function, there are other aspects of function prediction that make it more challenging. Protein function prediction methods that are based on protein family resources are often limited by the scope of the family resource and its ability to provide limited functional information. Moreover, function annotation of proteins in the databases are incomplete and biases in known function annotations or misannotations in the databases further affects our understanding of protein function space. Consequently, correct and highly specific function predictions can often be regarded as false-positives if they are currently annotated in the databases only in a generic manner. For example, for a protein that is annotated with only one general MFO term such as 'protein binding' in the databases, any function prediction method that would pre-

dict a more specific term in the MFO than 'protein binding' describing its function would be considered erroneous. However, the absence of the other MFO terms for the protein in the databases does not necessarily imply the absence of the related functions. This is popularly known as the Open World Assumption.

Thus, the performance of function prediction methods can vary with different benchmark categories depending on their application objectives. Also, assessment of the benchmark results of function prediction methods is not trivial and a range of different benchmarks, evaluation modes and metrics are required to reveal the strengths and weaknesses of a function prediction method. Although the recent independent assessment of methods for function prediction by CAFA organisers (Friedberg and Radivojac, 2016) have been extremely valuable for determining which function prediction approaches work well, they have also highlighted the associated challenges in providing reliable, accurate predictions. Interestingly, in both CAFA 1 and CAFA 2 assessments, methods relying purely on whole protein or domain homology were ranked among the top performing methods. This suggests that there is considerable signal in the domain sequence reflecting the protein's molecular function and the context in which it operates.

In this chapter, it was shown that sub-classifying protein domain superfamilies into functional families or FunFams using FunFHMMer helps in making precise function predictions for uncharacterised sequences by inheriting functional annotations from sequence relatives of the most closely related FunFam to the query sequence. Moreover, analysis of the CAFA 2 results indicated that prediction of function using domain sequence homology is a very powerful approach for functional annotation of uncharacterised sequences which is still competitive to current machine-learning approaches that combine multiple additional information such as protein interaction, gene expression and cellular localisation data.

Chapter 4

Using FunFams to explore functional diversity of CATH superfamilies and predict functional sites

4.1 Background

A large number of computational methods utilise protein family information to predict functions or functional sites (Ashkenazy *et al.*, 2010; del Sol Mesa *et al.*, 2003; Capra *et al.*, 2009; Lichtarge *et al.*, 1996; Innis *et al.*, 2004; Jones *et al.*, 2014). However, automated functional classification of protein superfamilies into families still remains a challenging task. The quality of protein families used for the prediction of functional annotations or functional sites can affect the performance of the prediction method to a great extent. As a result, protein resources or automated functional classification methods that provide functionally coherent protein families are of great importance to the protein function annotation community.

The functional families (or FunFams) in CATH-Gene3D generated by the FunFHMMer functional classification protocol (described in Chapter 2) have been found to be more functionally coherent than other domain-based resources (Das *et al.*, 2015b). The FunFHMMer classification protocol makes use of only sequence information to identify FunFams due to the lack of sufficient structural data to capture the functional diversity of a large number of protein superfamilies.

Few protein functional classification methods have utilised known structural or functional site information from the literature to identify families in selected superfamilies. One of the first family identification methods to do this was DASP (Cammer *et al.*, 2003). DASP (Deacon Active Site Profiler) identifies known functionally important residues of a protein superfamily and creates active site pro-

files for all known structures in the superfamily by extracting and concatenating the sequence fragments from a 10 Å radius around any of the known functionally important residues from N- to C-terminus. These active site profiles are then aligned and clustered to identify families. A similar approach was used by Nagao *et al.* (2010) to evaluate the utility of identifying families in seven diverse CATH enzyme superfamilies using known functional (i.e. both catalytic and ligand-binding) residue information. In this approach, a multiple structural alignment of representative enzymes in each superfamily was constructed and the functional residues were mapped to the structural alignment. Regions of the alignment containing the functional residues are then concatenated from N- to C-terminus and then clustered to identify families. The identified families were found to be more functionally pure than those generated by clustering of the full length domain alignments. The success of these family identification approaches using functional site information in selected superfamilies strongly suggests that the performance of functional classification protocols can be improved to a great extent by making use of structural data to focus on likely functional sites.

4.1.1 Protein functional sites

Protein functional sites are groups of amino acid residues that carry out the functional role of the protein. These include catalytic sites (for enzymes) and binding sites for chemical ligands, ions, other proteins and nucleic acids. Function annotation of proteins is incomplete without characterisation of their functional sites. Knowledge of such functionally important residues in proteins can guide targeted site-directed mutagenesis experiments, drug design and protein engineering.

4.1.1.1 Diversity of functional sites in superfamilies

One of the main challenges in prediction of functional sites using evolutionary information is the identification of homologous sequences that not only carry out

the same function as the query protein but also utilize the same functional sites. This is a non-trivial task considering the diversity of functional sites that have been reported even among homologous proteins that are closely-related (Gerlt *et al.*, 2005; Brown and Babbitt, 2014; Dessailly *et al.*, 2013; Furnham *et al.*, 2015).

Babbitt and co-workers (Brown and Babbitt, 2014) have reported that relatives of the functionally diverse SFLD superfamilies that catalyse different overall chemical reactions, generally share only one or more catalytic residues in common that are used for a common partial reaction. In another study (Wass *et al.*, 2011) on the conservation of ligand-binding sites in SCOP superfamilies, the authors reported that in most superfamilies, at least one binding site is generally highly conserved. For superfamilies containing only one binding-site, the site is generally conserved in most sequence relatives and for superfamilies with many binding sites, the conservation of sites can vary.

A recent large scale analysis of enzyme superfamilies in CATH has revealed that considerable sequence divergence can also occur in the active site region of proteins (Furnham *et al.*, 2015). In about two thirds of 101 enzyme superfamilies having experimental annotations on catalytic residues and reaction chemistry, dramatic changes in the catalytic machinery were reported. However, in 50% of these, at least one or two catalytic residues were found to be conserved among all superfamily relatives. Most of the superfamily diversity was observed to be associated with changes in substrate specificity.

Although the catalytic machinery is not completely conserved among homologous proteins, an in-depth study on the spatial diversity of functional sites in CATH superfamilies by Dessailly *et al.* (2013) showed that for a majority of the superfamilies, the spatial locations for catalytic sites are generally limited. However, members of large diverse superfamilies can show a considerable amount of functional plasticity and their relatives can exploit different sites for binding small-ligands or interacting with their protein partners. Furthermore, it was also

shown that relatives of such diverse superfamilies that are grouped within the same functional family have a greater tendency to exploit a common functional site (Dessailly *et al.*, 2013).

4.1.1.2 Methods for prediction of functional sites

Many computational approaches have been used to predict functional sites in proteins. These approaches can be sequence-based, structure-based or a combination of both. These include methods that exploit structural similarity and conservation information derived from alignments of homologous proteins. There are considerable differences in the properties of various types of functional sites which provide a basis for the prediction of these sites. Catalytic sites generally have limited exposure to solvent while binding sites have comparatively higher solvent accessibility. While both catalytic sites and ligand-binding sites are generally located in one of three largest clefts for a majority of proteins (Bartlett *et al.*, 2002b; Capra *et al.*, 2009), protein-protein binding sites or interfaces are relatively flat (Jones and Thornton, 1997). Moreover, catalytic site or ligand-binding sites are generally highly conserved. In contrast, protein-protein binding sites are difficult to predict from sequence conservation alone.

Sequence-based functional site prediction methods use sequence conservation information derived from multiple-sequence alignments of homologous proteins, however, some of these methods also use structural data onto which the conserved residue information is mapped. Evolutionary Trace (ET) (Lichtarge *et al.*, 1996), ConSurf (Ashkenazy *et al.*, 2010) and INTREPID (Sankararaman and Sjölander, 2008) are widely used functional site prediction programs that use phylogenetic analysis of homologous protein sequences to identify functionally important positions that are conserved at different levels of an evolutionary tree. Each of the methods provide a score indicating the likely functional importance of the site. While INTREPID uses sequence information only, ConSurf and ET meth-

ods map the predicted functional residues onto protein structures where they tend to cluster together in space forming three-dimensional residue clusters which are predicted as functional sites.

The Evolutionary Trace (ET) (Lichtarge *et al.*, 1996) method is a sequence-structure analysis technique developed to identify functionally and structurally important residues. ET method constructs a phylogenetic tree derived from an MSA for a query protein and its homologs. It then partitions the phylogenetic tree into distinct branches to identify functionally similar relatives and identifies highly conserved groups of residues within homologous proteins of each branch of the tree by correlating amino-acid variations in a multiple sequence alignment together with structural constraints. Figure 4.1 shows the detailed workflow of the ET method.

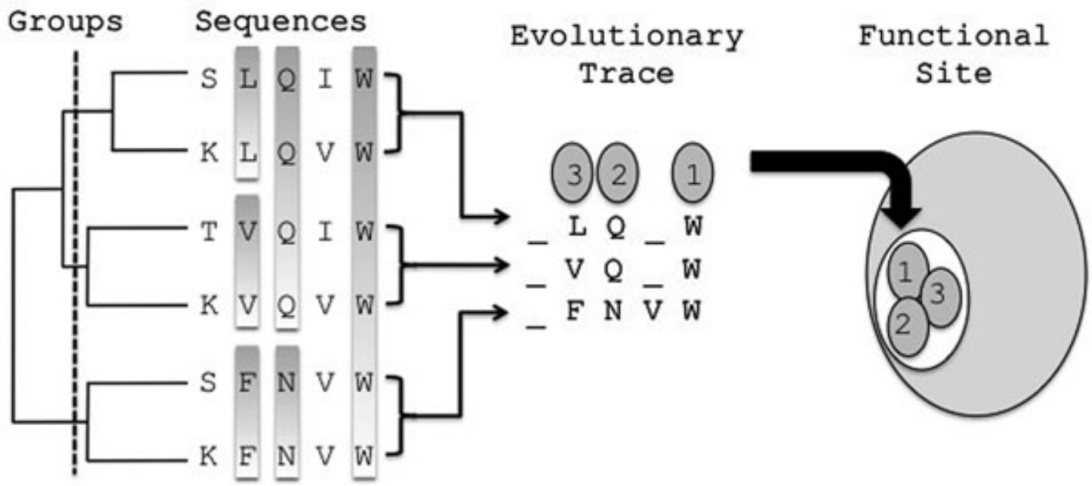


Figure 4.1: The Evolutionary Trace (ET) method for identifying functional site residues. First, the multiple sequence alignment of homologous sequences are divided into groups based on the phylogenetic tree and for each group, a consensus sequence is generated. The ET method extracts the relative evolutionary importance of the residues and assigns a rank to each residue accordingly. Residues with lower numbered ranks are considered to be more important than those with higher numbered ranks. The lower ranked residues (for example, 1, 2 and 3 in the given example) are then mapped onto a protein structure where clusters of residues indicate functional sites. Taken from Wilkins *et al.* (2012).

Predominantly structure-based methods either use the three-dimensional co-

ordinates of the amino acid residues to analyse the shape and structural features of a protein or the available structural data of ligands bound to proteins. Conventional methods generally identify pockets or clefts in proteins using geometry-based programs like SURFNET (Laskowski, 1995), LigSite (Hendlich *et al.*, 1997) and PASS (Brady Jr and Stouten, 2000) to predict sites where a ligand is likely to bind. For example, SURFNET identifies protein cavities by placing a sphere of appropriate size (between 1-4 Å) between two given atoms, such that they are on opposite sides of the sphere's surface and generating gap spheres (Figure 4.2). When all pairs of atoms are considered, the overlapping gap spheres delineates the shape and size of the gap regions or cavities in the protein. A number of methods have combined sequence conservation information and protein cleft or pocket prediction to improve prediction of catalytic and binding site residues such as SURFNET-ConSurf (Glaser *et al.*, 2006), LigSite^{csc} (Huang and Schroeder, 2006) and ConCavity (Capra *et al.*, 2009) which provide significant improvements over using the geometry based pocket or cleft prediction method alone. For example, the SURFNET-ConSurf (Glaser *et al.*, 2006) method first identifies clefts in a query protein for potential ligand binding sites and extracts residue conservation scores from the ConSurf-HSSP database (Glaser *et al.*, 2005) that estimates the evolutionary rate of each amino acid in a PDB structure. The SURFNET-ConSurf method then refines the predicted clefts by eliminating regions that are distant from the highly conserved residues. The resulting cleft regions are then predicted as ligand binding sites. Similarly, the ConCavity (Capra *et al.*, 2009) method integrates sequence conservation information directly into a pocket prediction method such as SURFNET or LigSite. It first projects a protein structure to a cubic grid and during the grid creation process, then it weights the pocket grids based on sequence conservation values. This helps it in identifying pocket regions that are potential catalytic or binding sites with better performance than considering the pocket prediction method alone.

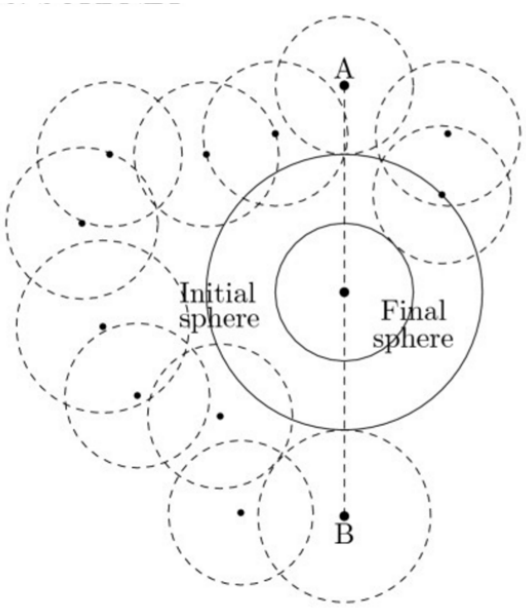


Figure 4.2: Identification of protein clefts by SURFNET. SURFNET identifies protein cavities by placing a sphere between two atoms and generating gap spheres. The spheres with maximal volume define the largest pocket. Taken from Huang and Schroeder (2006) under CC BY 2.0.

More recent structure-based methods for ligand-binding sites use available information on ligands bound to homologous structures based on the assumption that homologues are likely to utilise similar binding sites that bind the same or similar ligands (Wass *et al.*, 2011). These methods such as FINDSITE (Brylinski and Skolnick, 2008) and 3DLigandSite (Wass *et al.*, 2010) first identify ligand-bound homologous structures to the query sequence using threading and structural similarity searches respectively and then model the structure of the query protein. The ligand-bound structures are then aligned to the query model such that the ligands are superimposed onto the query model. The ligands are then clustered and the clusters are predicted as ligand-binding sites.

4.1.1.3 Assessment of functional site predictions

The performance of automated functional site prediction methods is generally measured using the Matthews Correlation Coefficient (MCC) and binding site distance test (BDT) (Schmidt *et al.*, 2011; Gallo Cassarino *et al.*, 2014). Both the

measures account for over and under predictions based on known functional site definitions from either the literature or curated databases.

For measuring the Matthews Correlation Coefficient (Matthews, 1975, MCC), the predicted functional residues are first classified in to the following categories based on known functional site definitions: (i) true positives (TP) or correctly predicted functional site residues, (ii) true negatives (TN) or correctly predicted non-functional site residues, (iii) false negatives (FN) or incorrectly under predicted functional site residues and (iv) false positives (FP) or incorrectly over predicted functional site residues. MCC is then measured by Equation 4.1 where the values of MCC can range from -1$\leq$MCC$\leq$1 and a score of -1 indicates incorrect prediction, 0 indicates random prediction and a score of 1 indicates a perfect or exact prediction of functional site residues.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}} \tag{4.1}$$

The binding site test (BDT) (Roche *et al.*, 2010) score takes in to account the three-dimensional coordinates of the predicted functional residues, and scores them according to the distance between the predicted and the known functional site residues using a distance threshold. The Euclidean distance (d_{ij}) between the C-α coordinates of each predicted residue i and each known functional site residue j is first calculated and the distance is converted to a S-score (S_{ij}) using Equation 4.2 where d_0 is the distance threshold.

$$S_{ij} = \frac{1}{1 + (\frac{d_{ij}}{d_0})^2} \tag{4.2}$$

The maximum S_{ij} score is determined for each predicted residue. The BDT score is then calculated as the sum of the maximum S_{ij} scores normalized by the greater value of the total number of predicted residues (N_p) and the number of known functional site residues (N_{func}) using Equation 4.3 where a score of 0

indicates a random prediction and 1 indicates a perfect prediction.

$$BDT = \frac{\sum_{i=1}^{N_p} \max(S_{ij})}{\max(N_p, N_{func})} \tag{4.3}$$

4.2 Aims and Objectives

This work examines the functional families or FunFams in CATH superfamilies and investigates how the quality of the FunFams can be improved. The FunFams are then used as a tool for exploring superfamily diversity in the CATH-Gene3D resource and identifying functional determinants in proteins using an in-depth manual analysis of the serine beta-lactamases. This analysis has been published in:

Lee, D., Das, S., Dawson, N. L., Dobrijevic, D., Ward, J. and Orengo, C. A. (2015). Novel computational protocols for functionally classifying and characterising serine beta-lactamases, *PLoS Computational Biology*, 12(6), e1004926.

Based on the utility of the FunFams in identifying functional site residues, the recent development of the FunSite method is then described which predicts active site and ligand-binding residues by exploiting evolutionary information in FunFam alignments and structural data available for the FunFam sequence relatives.

4.3 Analysing and improving the quality of CATH FunFams

4.3.1 CATH (v4.0) statistics

The CATH (version 4.0) database contains over 235,000 structural domains comprising 1375 unique folds that have been sub-classified into 2735 superfamilies. The corresponding Gene3D (v12.0) database contains over 25 million sequence domain predictions. The GeMMA clustering and FunFHMMer classification al-

Table 4.1: CATH (v4.0) statistics showing the total number of Gene3D domain sequences and CATH structural domains in the CATH-Gene3D resource and the number of Gene3D domain sequences and CATH structural domains that could be assigned to FunFams.

Sequence/structural domains	Total no. in CATH-Gene3D v4.0	Total no. assigned to CATH FunFams v4.0
Gene3D domain sequences	$\sim$ 25.6 million	4.56 million seed sequences + 7.14 million non-seed sequences = $\sim$ 11.7 million (45.73%)
CATH structural domains	235,000	179,826 (76.52%)

gorithms were used for classifying the CATH-Gene3D superfamilies into functional families or FunFams. This functional classification pipeline pre-clustered all superfamily sequences at 90% sequence identity into S90 clusters and only processed those S90 clusters that had at least one sequence with high-quality GO annotations. As a result, only 4.56 million sequences ($\sim$18%) out of over 25.6 million Gene3D sequences were used as seed sequences for the functional classification pipeline which resulted in generation of 110,439 FunFams in 2735 CATH-Gene3D superfamilies.

After the FunFams were identified for all the superfamilies, $\sim$20.5 million Gene3D domain sequences that were not used as seed sequences (i.e. non-seed) and the 235,00 CATH structural domains were scanned against the FunFam HMM models and were assigned to the FunFams if they exceeded the inclusion threshold score of the respective FunFam HMM model. Almost 76.5% of the CATH structural domains were assigned to the FunFams. However, only over 7.14 million (34%) out of 20.5 million non-seed Gene3D sequences achieved the inclusion threshold of any of the FunFams. This results in a total of over 11.7 million Gene3D sequences (45.73%) and 179,826 CATH structural domains (76.52%) comprising the 110,439 FunFams (Table 4.1).

4.3.2 Analysis of FunFams for improving their quality

Low percentage of Gene3D sequences assigned to FunFams

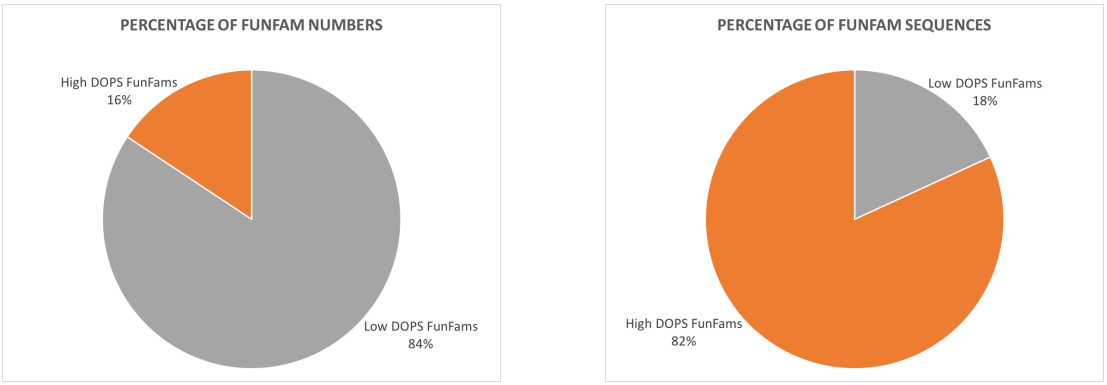
The low percentage of Gene3D domain sequences that can be assigned to CATH FunFams may be either due to one of the following reasons:

(i) *Annotated seed sequences do not represent the sequence diversity of the whole superfamily* - The seed sequences of a superfamily i.e. those sequences that have at least one sequence relative with $\geq$ 90% sequence identity having at least one high-quality GO annotation do not provide sufficient representation of the sequence diversity of the superfamily. Thus, it may be possible to assign more Gene3D sequences to the FunFams by including any S90 clusters in the function classification pipeline that have at least one sequence annotated with IEA (i.e. electronically annotated) GO annotations, as well as high-quality GO annotations as the overall quality of electronic annotations has been reported to be reliable (Škunca *et al.*, 2012). However, this would be challenging for the highly populated superfamilies in CATH since the associated tools in the pipeline would have to deal with increased amounts of sequence data. Furthermore, in time many of these less well-annotated sequences will receive high-quality GO annotations, and seed FunFams.

(ii) *Conservative FunFam model inclusion thresholds* - The inclusion threshold of the FunFams are very conservative. However, it is not trivial to define the inclusion threshold of a FunFam as it will directly affect the functional purity of the FunFam. The inclusion thresholds used currently are deliberately conservative as FunFams are primarily utilised for functional annotation (see Chapter 3). However, in other contexts e.g. to provide assignments for distant sequence relatives such as in structural modelling, it may be valuable to lower the FunFam model inclusion thresholds.

Large number of FunFams with low information content

Only 17,326 FunFams (16%) are highly informative (Figure 4.3a) i.e. the FunFam alignments comprise of evolutionary distant relatives that provide more discriminating conservation scores during conservation analysis and have high DOPS scores ($\geq$ 70, described in Section 1.2.1.2 in Chapter 1). However, 82% ($\sim$ 9.6 million) of the FunFam sequences can be mapped to these highly informative FunFams (Figure 4.3b). The remaining less informative FunFams having low ($<$ 70) DOPS scores, comprising about 18% of the FunFam sequences, either contain a small number of highly similar sequences or a single sequence and hence, cannot be used for conservation analyses.



(a) **(b)**

Figure 4.3: (a) Piechart showing the percentage of CATH FunFams with high information content i.e. with high DOPS scores ($\geq$ 70) and low information content i.e. with low DOPS scores ($<$ 70). **(b)** Piechart showing the percentage of CATH FunFam sequences that are assigned to FunFams with high information content (high DOPS scores) and those that are assigned to FunFams with low information content (low DOPS scores).

A large proportion of these less informative FunFams may have been generated by FunFHMMer due to one or both of the following reasons:

(i) *Less discriminatory conservation scores in less informative cluster alignments* - As the FunFHMMer strategy analyses the functional coherence of all parent nodes in the GeMMA hierarchical clustering tree in a bottom-up manner, the parent sequence cluster alignments at the bottom of the hierarchical tree are

likely to share very high sequence identities i.e. they are likely to be less diverse or less informative alignments. Consequently, analysis of pairs of these cluster MSAs may sometimes lead to false prediction of specificity-determining positions (SDPs) as a result of less discriminatory conservation scores which may lead the protocol to infer two functionally coherent MSAs as functionally not coherent. This would prevent merging of functionally coherent MSAs.

(ii) *Heuristics in GeMMA* - The structure of the GeMMA clustering tree for a superfamily which guides and limits the pairwise sequence cluster comparisons of the FunFHMMer protocol may also cause problems. This is because the heuristics integrated in the GeMMA algorithm may have potential impacts on the clustering tree (described in Section 4.1) that is used by the FunFHMMer pipeline.

Although these two limitations were considered during the development of the FunFHMMer protocol by using different criteria to infer functional coherence for highly informative and less informative sequence clusters (see Section 2.3.2 in Chapter 2) and modification of the GeMMA tree (see Section 2.3.3 in Chapter 2), there is a lot of scope for improvement in the quality of the FunFams. Moreover, it may be beneficial to take into consideration the sequence diversity and known functional diversity of a superfamily when calculating the functional coherence of sequence clusters of a particular superfamily as different superfamilies have undergone varying amounts of divergence in terms of sequence and function.

4.4 Exploring superfamily diversity using FunFams

The functional classification of the CATH superfamilies into FunFams were used for exploring superfamily diversity in the CATH-Gene3D resource. The top 200 superfamilies in CATH that have the highest number of FunFams (each of which have 100 FunFams or more), comprising $\sim$7% of the superfamilies and accounting for $\sim$65% of CATH structural domains (Figure 4.4) were analysed. The analysis showed that sequence changes in these top 200 superfamilies (shown

as pink or blue circles in Figures 4.6, 4.5 and 4.7) is associated with large amounts of diversity in structure, function and protein context. In contrast, the remaining ~93% of the CATH superfamilies (shown as grey circles in Figures 4.5, 4.6 and 4.7) appear to have structurally and functionally conserved relatives. In fact, 156 folds in CATH contain only one superfamily where all sequence relatives have a single FunFam and in total, 360 superfamilies have only one FunFam. Moreover, 40% of the CATH superfamilies have less than 5 FunFams.

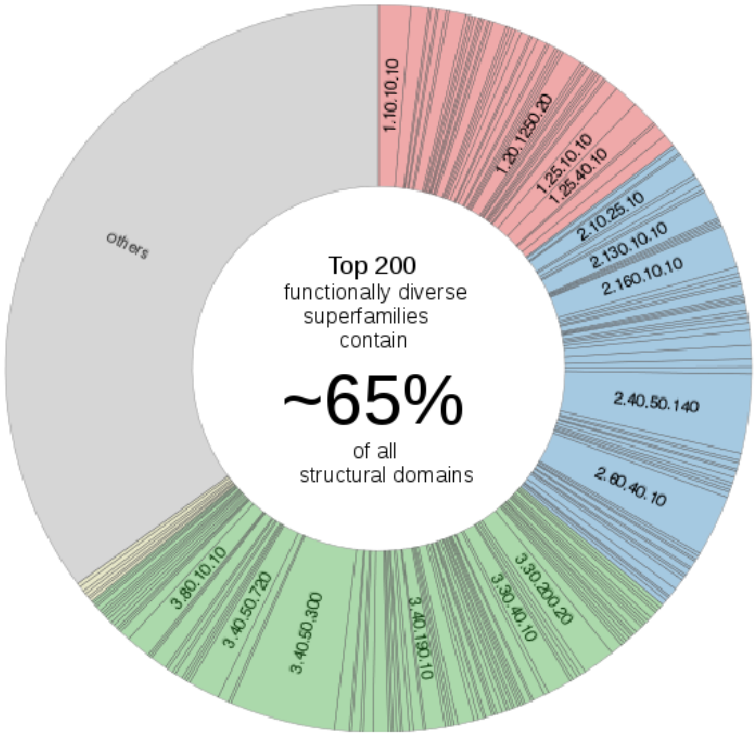
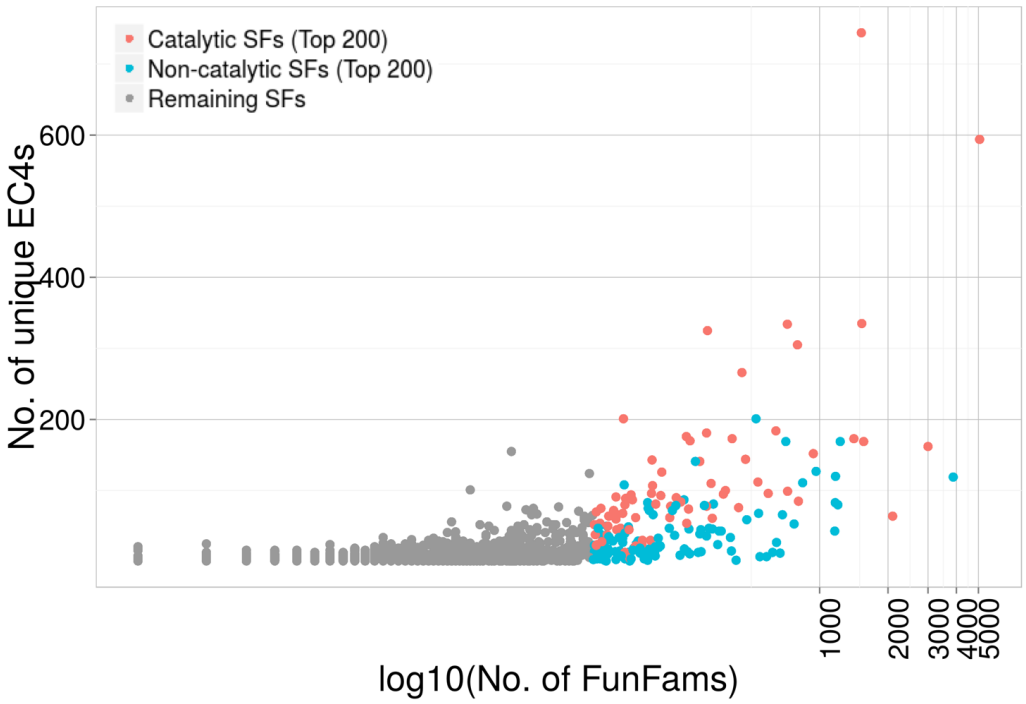


Figure 4.4: Graph showing the relative size of the top 200 CATH superfamilies ranked by the number of FunFams. The superfamilies are coloured according to their Class code - Classes 1, 2, 3 and 4 are coloured in red, blue, green and yellow respectively. The rest of the superfamilies are shown in grey.

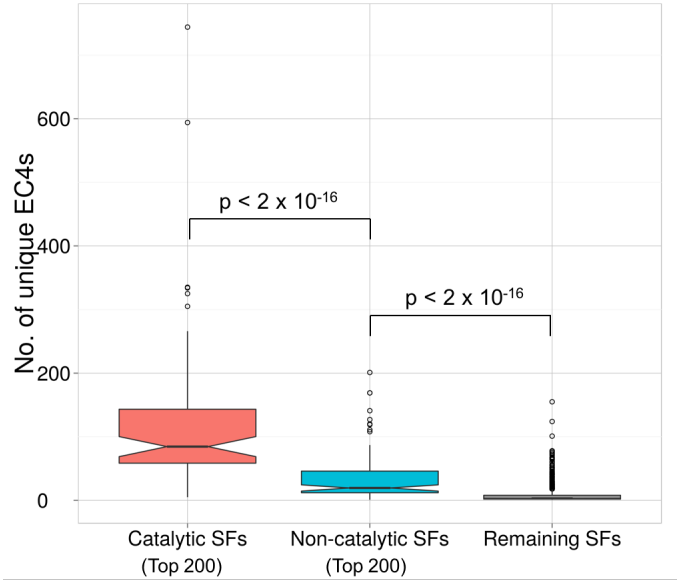
Recently, Furnham *et al.* (2015) have reported 384 superfamilies in CATH that contain protein domains that are solely responsible for enzymatic catalysis i.e. those containing the majority of the catalytic residues from CSA (Porter *et al.*, 2004). These superfamilies are referred to as catalytic domain superfamilies, and make up 36% of the top 200 superfamilies that have the highest number of FunFams. The remaining 64% of the top 200 superfamilies are referred to as

non-catalytic domain superfamilies.

Figure 4.5a shows the diversity of EC numbers (at the fourth level; EC4s) for 1983 (72.5%) CATH superfamilies that contain at least one enzyme-associated protein domain. The catalytic and non-catalytic domain superfamilies that feature among the top 200 superfamilies ranked by the number of FunFams, are shown in the figure as pink and blue circles respectively. The remaining superfamilies are shown in grey. The top 200 superfamilies are found to show significant diversity in EC numbers compared to the rest of the superfamilies (Figure 4.5b). Moreover, the catalytic domain superfamilies among the top 200 superfamilies that have the highest number of FunFams generally seem to show significantly more diversity in EC numbers than the non-catalytic domain superfamilies (Figure 4.5b).



(a)

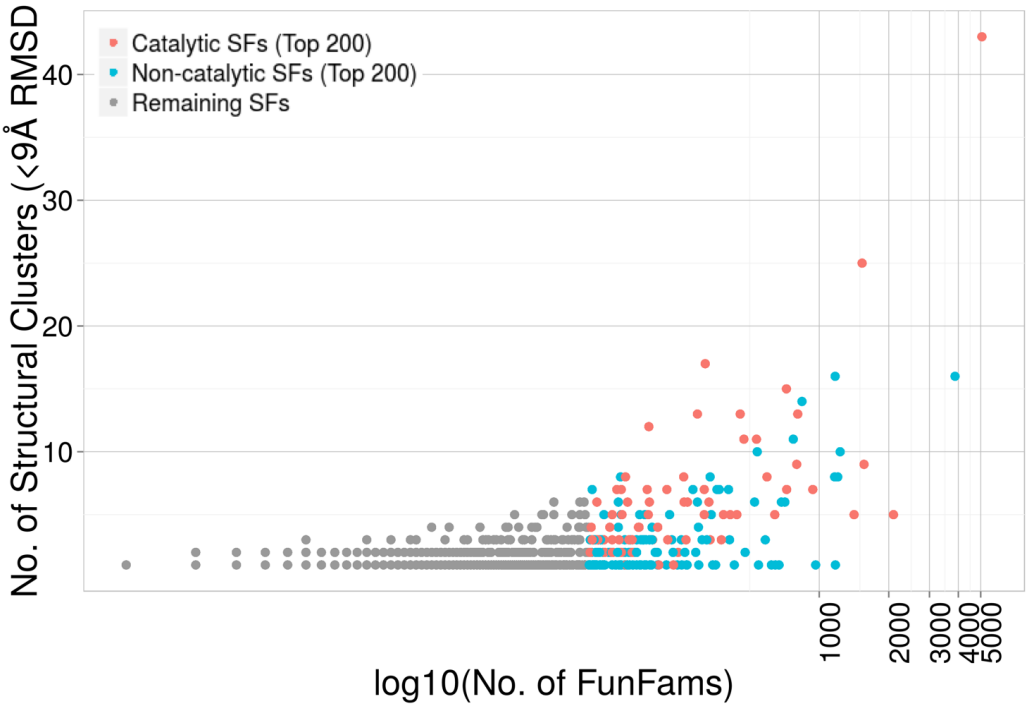


(b)

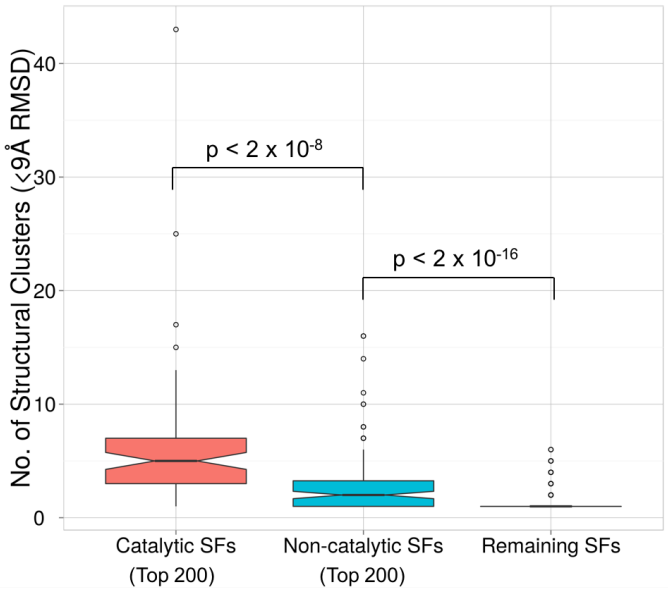
Figure 4.5: EC diversity in CATH superfamilies. (a) Correlation of the diversity of EC numbers at the fourth level (EC4s) with the number of FunFams in CATH enzyme superfamilies (plotted in the logarithmic scale). The EC4 diversity of a superfamily is shown by the number of unique EC4s. Each circle in the figures represents a CATH superfamily. The top 200 CATH superfamilies ranked by the number of FunFams are either coloured in pink (showing catalytic domain superfamilies) or blue (showing non-catalytic domain superfamilies). The remaining superfamilies are coloured grey. (b) Box plot of the diversity of EC4 numbers are shown. p-values indicated in the plot were calculated using the Wilcoxon Rank-Sum tests.

Similarly, the structural and multi-domain architecture (MDA) diversity of all CATH superfamilies is illustrated in Figures 4.6 and 4.7. The structural diversity of a superfamily is shown by the number of distinct Structurally Similar Groups (SSGs) in which relatives superpose with $<$ 9Å RMSD (Cuff *et al.*, 2009) and the MDA diversity is shown by the number of different MDAs containing one or more superfamily domains. The top 200 superfamilies having the highest number of FunFams are found to show significant diversity in structure and multi-domain architecture compared to the remaining superfamilies (Figures 4.6b and 4.7b). Also, the catalytic domain superfamilies among the top 200 superfamilies were found to be significantly more structurally diverse than the non-catalytic domain superfamilies (Figure 4.6b).

In summary, the functional classification of CATH superfamilies into FunFams was analysed to explore the superfamily diversities captured by the FunFams. It was shown that FunFams capture well the structural, functional and domain architecture diversity in superfamilies. Moreover, the top 200 superfamilies in CATH ranked by the number of FunFams accounts for nearly two thirds of all CATH structural domains and show significant diversity in terms of function, structure and multi-domain architectures compared to the rest of the superfamilies. Moreover, the subset of the top 200 superfamilies that contain domains that are responsible for enzyme catalysis tend to me more structurally and functionally diverse than the other superfamilies among the top 200 superfamilies.



(a)



(b)

Figure 4.6: Structural diversity in CATH superfamilies. (a) Correlation of the structural diversity with the number of FunFams in CATH superfamilies (plotted in the logarithmic scale). The structural diversity of a superfamily is shown by the number of distinct Structurally Similar Groups (SSGs) in which relatives superpose with $<$ 9Å RMSD. The color scheme is same as in Figure 4.5. (b) Box plot of the number of distinct SSGs are shown. p-values indicated in the plot were calculated using the Wilcoxon Rank-Sum tests.

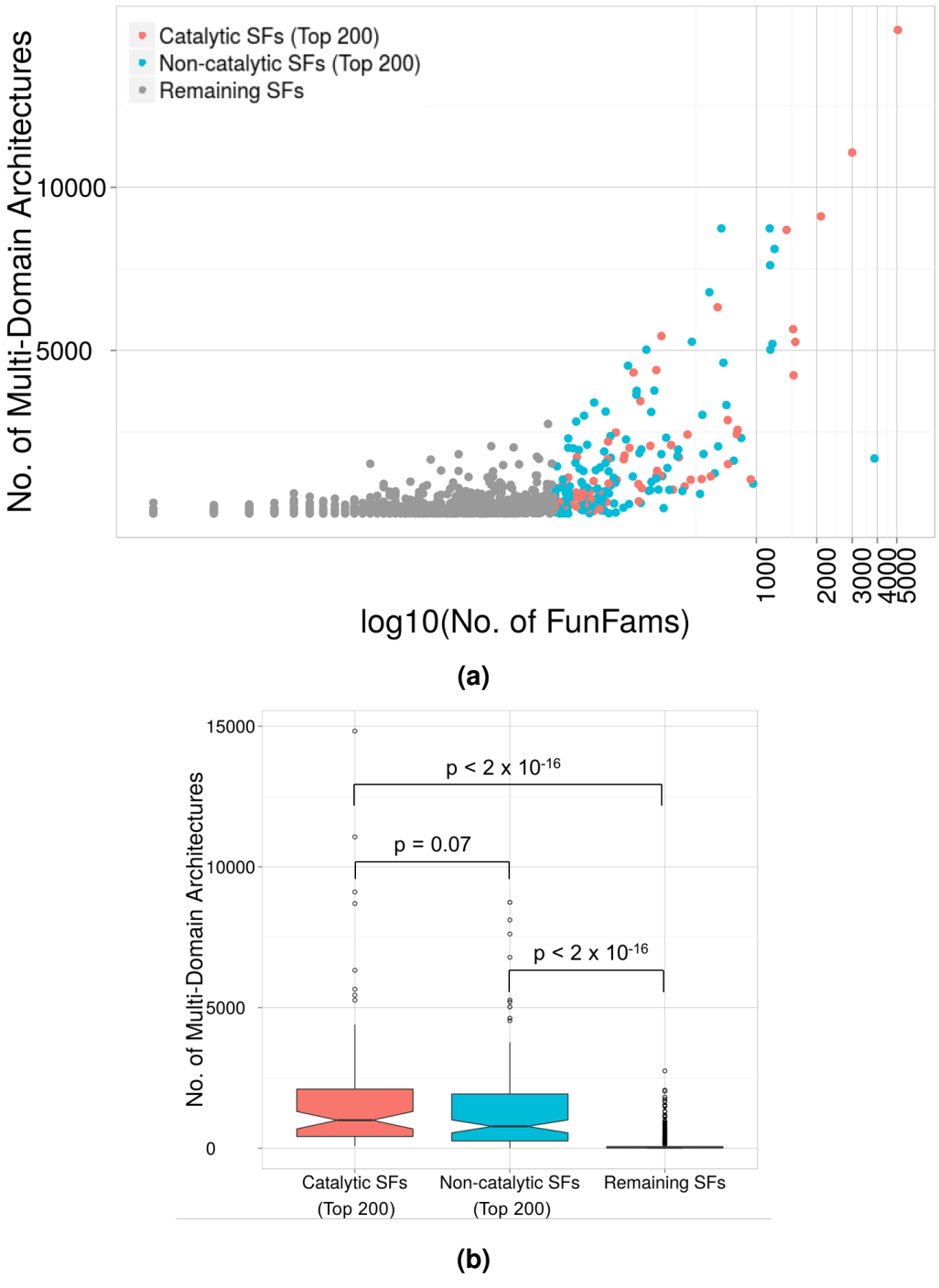


Figure 4.7: Multi-domain architecture (MDA) diversity in CATH superfamilies. (a) Correlation of the MDA diversity with the number of FunFams in CATH superfamilies (plotted in the logarithmic scale). The MDA diversity of a superfamily is shown by the number of different MDAs containing one or more superfamily domains. The color scheme is same as in Figure 4.5. (b) Box plot of the number of different MDAs are shown. p-values indicated in the plot were calculated using the Wilcoxon Rank-Sum tests.

4.4.1 Network visualisation of FunFam relationships

Many highly-populated superfamilies in CATH have considerable structural and functional diversity. A fast and easy approach in capturing this diversity and gaining useful insights about function evolution in superfamilies from a large-scale perspective is by the use of sequence similarity networks (Atkinson *et al.*, 2009). Recently, a large number of superfamily studies have used sequence similarity network-based approaches based on full-length sequences (Brown and Babbitt, 2014).

For all CATH domain superfamilies having two or more FunFams, superfamily networks have been constructed using Cytoscape (Smoot *et al.*, 2011) in which FunFams are represented by nodes and the edge distances correspond to the sequence similarity between the FunFam HMMs assessed using Profile Comparer (PRC) (Madera, 2008). The networks are visualised in the prefuse force-directed layout with edges weighted by the PRC score. These networks provide a comprehensive summary of domain sequence, structure and function relationships in a CATH superfamily. For example, Figures 4.8 and 4.9 shows two networks for the structurally and functionally diverse HUP superfamily that are useful for understanding how function has been modulated by sequence or structure changes between the FunFams. In Figure 4.8, the nodes are coloured by the EC number of constituent sequences. It can be seen from the figure that most of the time, FunFams having the same EC number (i.e same colour) cluster together. Figure 4.9 shows a reduced network of FunFams of high information content in the HUP superfamily and structural diversity among the FunFams of high information content in the superfamily. These networks can aid in the identification of potential novel targets for experimental characterization.

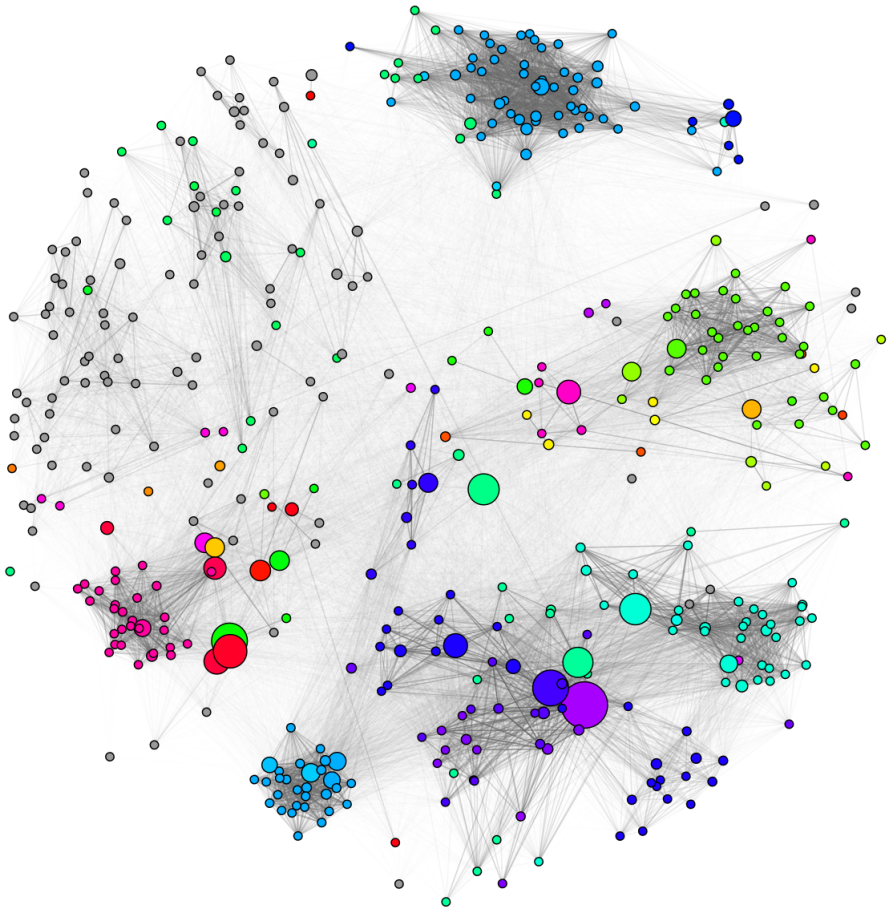


Figure 4.8: Visualisation of functional diversity in the HUP superfamily using Cytoscape (Kohl *et al.*, 2011) networks. The nodes in the network represent FunFams and the edges represent sequence similarities between the FunFam HMMs calculated using PRC (Madera, 2008). The size of the nodes (FunFams) reflect their population in number of sequences and the nodes are linked by edges if the similarity of their HMMs is above a PRC score of 10. This network highlights the functional diversity of the HUP superfamily where all nodes are coloured according to the EC numbers of their constituent sequences and grey nodes indicate those without any EC annotation (including non-enzymes).

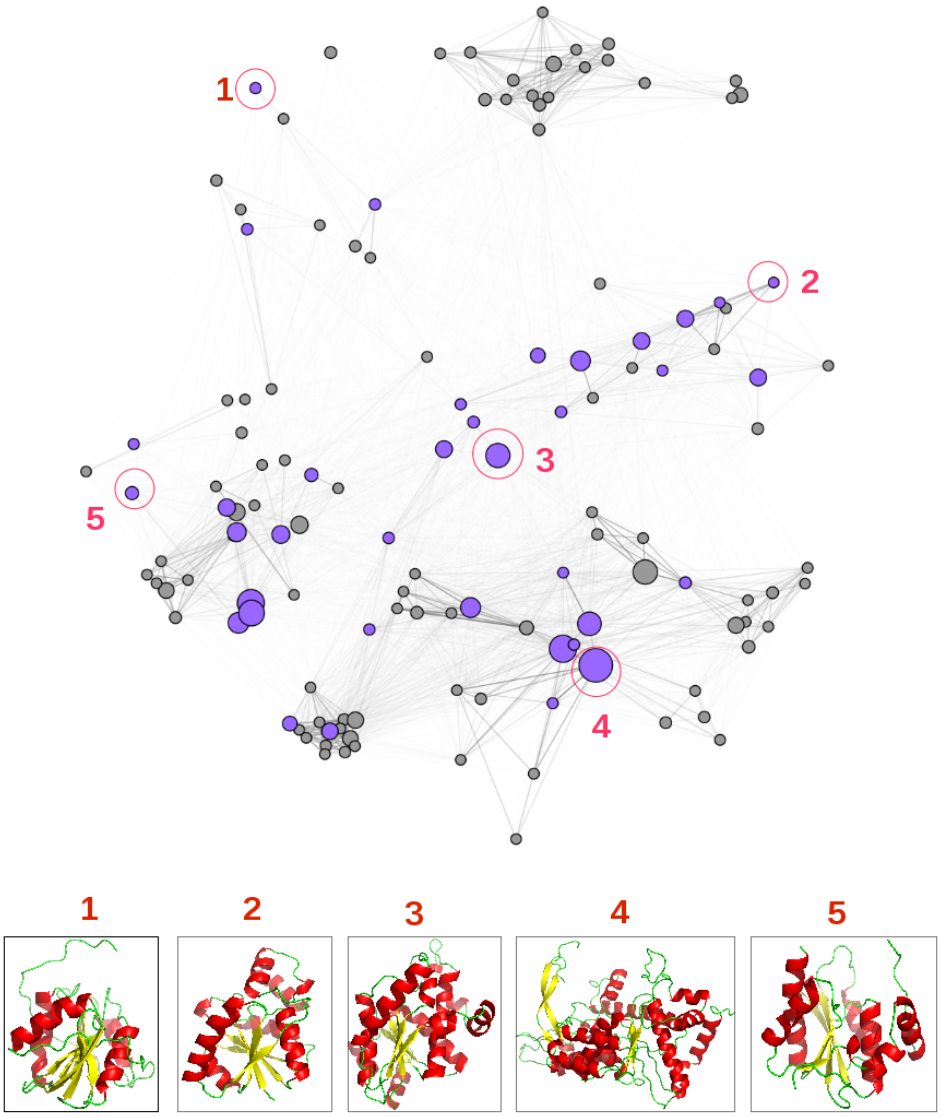


Figure 4.9: Visualisation of structural diversity in the HUP superfamily using Cytoscape (Kohl *et al.*, 2011) networks. The nodes represent FunFams and edges represent sequence similarities between the FunFam HMMs as described in Figure 4.8. This network shows the available structure data among the FunFams with high information content in the HUP superfamily. The purple coloured nodes indicate FunFams with known structure and the grey nodes indicate FunFams without any known structure. Structural representatives of selected FunFams (encircled and numbered in red) are shown at the bottom of the figure to highlight the structural diversity of the superfamily.

4.5 Identification of functional sites using FunFams

As well as using FunFams for predicting functional annotations i.e. GO or EC terms for proteins (see Chapter 3), FunFams can also be used for predicting protein functional residues since the conserved sites in FunFams are known to be highly enriched in catalytic site residues (see Section 2.3.7, Chapter 2). Thus, information in the conserved sites in FunFams can be used to determine the location of residue sites essential for protein functions, thereby, providing insights into the mechanism of the protein and can be very useful for identification of functional determinants. To manually assess this, a case study was performed on the identification of functional determinants (FDs) in serine beta-lactamases (proteins that degrade beta-lactam antibiotics that result in antibiotic resistance) using the FunFams.

4.5.1 A case study on identification of protein functional determinants using FunFams

The serine beta-lactamases were chosen for this study as they have been very well-studied as there is significant structure and sequence data available. They were also chosen as our collaborators in the Midwest Center for Structural Genomics (MCSG) were interested in target selection of different variants of the beta-lactamases. The aim was to identify relatives with different antibiotic resistance properties.

4.5.1.1 Serine beta-lactamases

Beta-lactamases (or β-lactamases, EC 3.5.2.6) are enzymes produced by some Gram-negative and Gram-positive bacteria that provide resistance to beta-lactam antibiotics such as penicillins, cephalosporins and carbapenems by hydrolysing the amide bond of the core beta-lactam ring (Figure 4.10a), thereby inactivat-

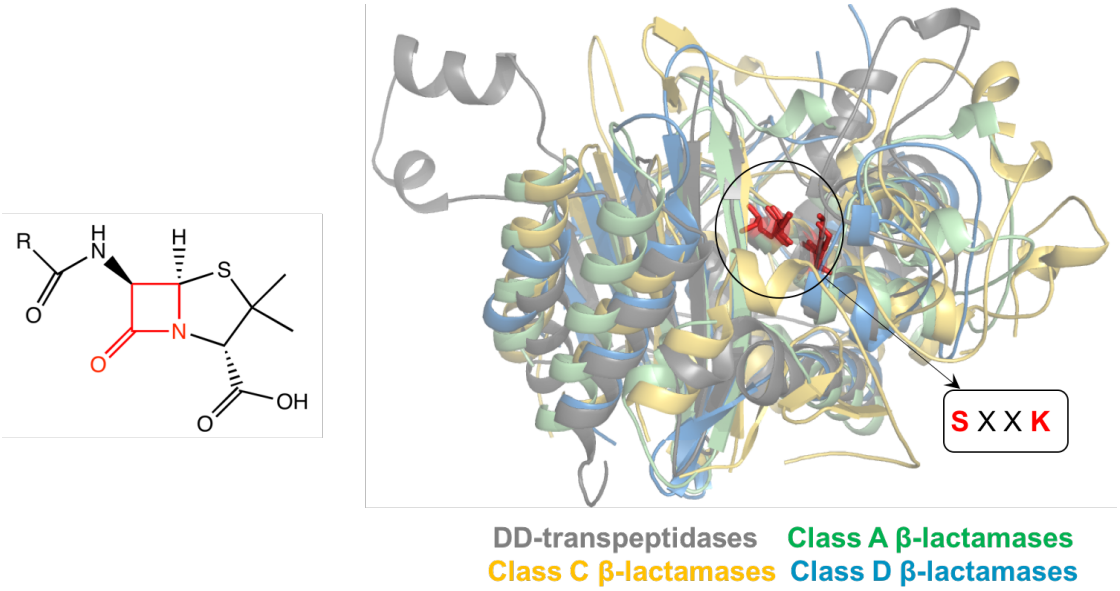


Figure 4.10: (a) Core structure of penicillins (a group of beta-lactam antibiotics) where 'R' is the variable group. The beta-lactam ring is highlighted in red. **(b)** Multiple structural alignment of representative structures of Classes A, C and D serine beta-lactamases and DD-peptidases. The catalytic serine (S70) and lysine (K73) pair, conserved in all the enzymes, is also shown.

ing it. The beta-lactam antibiotics act by inhibiting DD-peptidase (also known as DD-transpeptidase) enzymes that are responsible for the cross-linking of peptidoglycan units within the bacterial cell wall. The bacteria becomes sensitive to a variety of environmental stresses in the absence of a cell wall and this results in cell lysis. Excessive use of antibiotics worldwide have resulted in the emergence of increased beta-lactamase mediated resistance which poses a serious threat to modern medicine.

About 2 billion years ago, fungi evolved the ability to synthesize beta-lactam antibiotics that bind irreversibly to DD-peptidases, thus, inhibiting their activity. In response to this, it is presumed that some bacterial DD-peptidases have evolved into beta-lactamases that can break open the beta-lactam ring in an antibiotic, thereby inactivating it (Hall and Barlow, 2004). Serine beta-lactamases comprise three classes (Classes A, C and D) of beta-lactamases that utilise a catalytic serine residue in the breakdown of antibiotics and share the same structural fold

and CATH superfamily (3.40.710.10, DD-peptidase/Serine Beta-Lactamase superfamily) as the DD-peptidases. In the DD-peptidase/Serine Beta-Lactamase superfamily, although the serine beta-lactamases tend to have lower structural similarity with the DD-peptidases than with each other, the structural core is conserved across the whole superfamily (Figure 4.10b). In particular, a catalytic serine (S70) and lysine (K73) pair in the active site of DD-peptidases and serine beta-lactamases, superpose well. The residue positions of serine beta-lactamases are generally referenced by their corresponding structural equivalent residue in PDB 1SHV. This is referred to as the Ambler numbering system.

The three classes (Classes A, C and D) of serine beta-lactamases have evolved different solutions to degrade beta-lactam substrates and the major difference between the three classes is that they employ different implementations of the same general mechanism of action. This involves acylation followed by deacylation of the beta-lactam by the enzymes (Fenollar-Ferrer *et al.*, 2008). The general mechanism for all three Classes is as follows: (i) During acylation, the same structurally-equivalent catalytic serine is activated using a base which performs a nucleophilic attack on the beta-lactam ring, breaking the amide bond and forming an acyl-enzyme intermediate with the antibiotic. (ii) During deacylation, the acyl-enzyme intermediate undergoes hydrolysis. A water molecule is activated by a base for nucleophilic attack on the carbonyl bond of the acylenzyme, which releases the serine and the hydrolysis end-product from the acyl-enzyme intermediate and the enzyme is regenerated.

Although the mechanisms of the three Classes have been investigated extensively, they are still not completely clear as for each Class several hypotheses and lines of evidence exist. According to the current hypotheses and the MACiE (Holliday *et al.*, 2007) database, for Class A, a glutamate (E166) residue acts as the base for both acylation and deacylation steps; for Class C, a tyrosine (Y130) residue acts as the base for acylation and a lysine (K70) residue acts as the

base for deacylation and for Class D, a carboxylated lysine (K70) residue acts as the base for both acylation and deacylation steps. All the above residues lie close (within 5 Å) to the catalytic serine. Thus, there are differences in the bases that activate the catalytic serine during acylation and those that activate the water molecule that performs the hydrolysis of the acyl-enzyme intermediate. In addition, there are differences in the residue types hydrogen bonding to and protonating the amide nitrogen atom and other residue differences which have not yet been attributed a functional role.

4.5.1.2 Classification of serine beta-lactamase classes by FunFams

Conventional sequence comparison methods such as BLAST can be used to recognise closely related sequences i.e. greater than 60% identity for each Class of serine beta-lactamases. However, since distant relatives in each Class share less than 30% sequence identity with each other, more sensitive techniques are needed to completely distinguish the Classes.

The FunFHMMer protocol (described in Section 2.3.2 in Chapter 2) was used to sub-classify the DD-peptidase/Serine Beta-Lactamase superfamily into distinct functional families or FunFams. Manual inspection of the UniProtKB descriptions of the serine beta-lactamases confirmed that three FunFams of the superfamily captured well the three serine beta-lactamase classes A, C and D respectively. Small manual adjustments were made to the FunFams that resulted in complete agreement between the FunFam classification and beta-lactamase classes. For each serine beta-lactamase Class FunFam, the experimental annotations given in UniProtKB were inspected and any sequences having non beta-lactamase annotations e.g. having a DD-peptidase annotation were removed. These comprised fewer than 2% of sequences within each FunFam. It was assessed whether the FunFHMMer predictions of conserved residues in the FunFams of each of the three classes captured the residue differences in the active sites reported in the

literature, and whether FunFHMMer could reveal additional sites distinguishing these classes.

4.5.1.3 Functional determinants (FDs) identified using CATH FunFams

A three way multiple structural alignment was first built by selecting non-redundant sequences (at 60% sequence identity) with known structure, from each beta-lactamase Class FunFam and constructing an alignment by performing successive pairwise structure alignments using SSAP (Taylor and Orengo, 1989) against the representative that best matches all other representatives. After this, *hmmbuild* from the HMMER3 (Eddy, 2009) software was used to create an HMM for the structure alignment. Sequence relatives from the Class A, C and D FunFams were then aligned to the HMM using the *hmmalign* command from the HMMER3 software. GroupSim (Capra and Singh, 2008) was then applied to the resulting structure-based sequence alignment to find functional determinants (FDs) i.e. residues conserved in one class but not conserved, or conserved in a different way, in another class.

A structure-based sequence alignment was used instead of a multiple sequence alignment of the FunFams as the sequence alignment of the Class A and C FunFams together using the default parameters or *mafft-profile* option of the MAFFT algorithm did not align two functionally important and structurally-equivalent residues at Ambler position 130 i.e. the serine (S130) in Class A and tyrosine (Y130) in Class C (see Figure 4.11). This highlights the value of using structural data, where available, to guide an alignment of remote homologues.

The residue positions in the three serine beta-lactamase Classes that were identified by the FunFams as functional determinants are shown in Figure 4.11) using sequence logos of the three Classes. The functional determinants are also listed in Table 4.2. Two of these functional determinants (Ambler residue positions 130 and 166) are already known in the literature for contributing to the

implementation of the mechanism for different Classes. These two functional determinants along with a predicted functional determinant, Ambler position 211, that is likely to have an important functional role are the discussed in details in this section.

Table 4.2: Functional determinants that distinguish the three serine beta-lactamase classes predicted by applying GroupSim (Capra and Singh, 2008) to a structure-based sequence alignment of the A, C and D classes of serine beta-lactamase FunFams in the CATH superfamily 3.40.710.10. The predicted functional determinants identified in each FunFam are listed in the table along with their proportion of incidences in a FunFam. For simplicity, only residues having proportion greater than 0.1 are listed. Known catalytic site residue positions in any of the Classes are marked by asterisk (*).

Ambler residue	Distance (Å) from Serine 70	Residue in Class A FunFam	Residue in Class C FunFam	Residue in Class D FunFam
130*	4.98	S (0.98)	Y (1.00)	S (1.00)
131	8.1	D (1.00)	A (0.66) S (0.34)	V (0.48) T (0.33) A (0.15)
133	10.36	T (0.64) S (0.16)	S (0.48) P(0.26) A (0.18)	W (1.00)
134	10.58	A (0.98)	S (0.93)	V (0.77) Y (0.15)
136	10.18	N (0.98)	G (1.00)	Q (0.69) E (0.29)
157	14.96	D (0.96)	L (0.92)	Y (1.00)
166*	4.84	E (1.00)	A (0.78)	W (0.96)
179	9.23	D (0.95)	V (0.85) I (0.11)	L (0.98)
185	14.92	A (0.79) S (0.15)	D (1.00)	E (0.98)
211	11.21	M (0.69) L (0.30)	E (0.99)	M (1.00)
245	7.75	N (0.59) G (0.34)	S(0.59) A(0.35)	W (1.00)
246	9.08	D (0.61) I (0.35)	Y (1.00)	F (0.63) W (0.33)

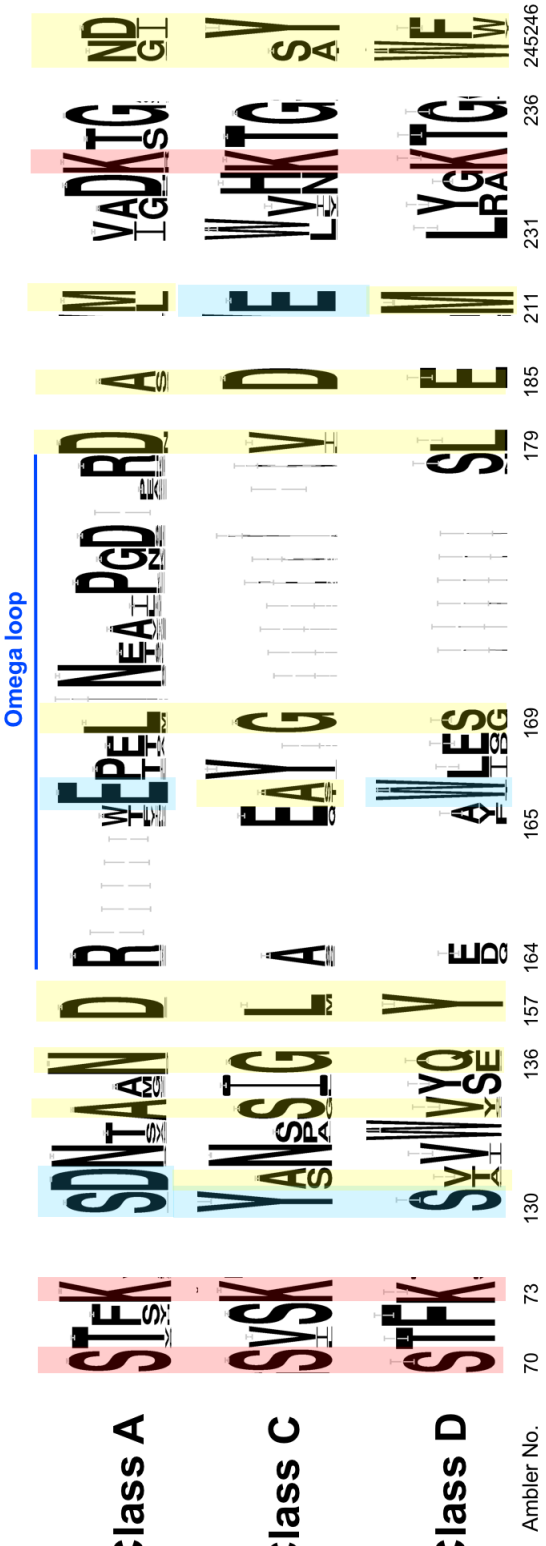


Figure 4.11: Sequence logo of the three-way structure-based sequence alignment of three classes (A, C and D) of serine beta-lactamase FunFams in the CATH superfamily 3.40.710.10. The Ambler numbering scheme is used to label the residue positions. FunFHMMer-identified conserved positions, predicted to be functional determinants, are coloured and the height of a character indicates the degree of conservation. The catalytic residues (S70, K73 and K234), all of which are predicted by FunFHMMer, are shown in red. Other FunFHMMer predicted residues which are also cited in the literature (including MACiE) are shown in blue, whilst those in yellow are predicted but not yet cited in the literature. Taken from Lee *et al.* (2016).

Ambler residue 130

A well-known position which differentiates between the three classes, and identified by FunFams, is the Ambler residue 130, which is a catalytic serine in Class A and D protonating the amide nitrogen atom of the beta-lactam ring after formation of the tetrahedral intermediate. By contrast, Class C has a catalytic tyrosine at position 130, which is also implicated in activating the hydrolytic water during the deacylation step. Figure 4.12 shows the Ambler residue 130 in Class A and Class C serine beta-lactamases. Ambler residue 131 is also identified as having a functional role by FunFams, in Class A (aspartate). Mutation studies reported in the literature have suggested that this residue is important for maintaining the enzyme activity. The corresponding residues in the other two classes are different but also conserved, although to a lesser extent, and so may also play a functional role.

Ambler residue 166

Another well-known function determinant identified by FunFams is the Ambler residue 166 which is a catalytic glutamate in Class A, activating the hydrolytic water for the acylation and deacylation steps. Different residues are found at this position in the other two classes - alanine in Class C and tryptophan in Class D. The tryptophan in Class D, W166, is known to be involved the hydrogen bonding network near the catalytic serine and lysine, however, the exact role of the alanine residue in Class C is not yet known. The catalytic glutamate, E166, in Class A beta-lactamases lies in the 'omega-loop' region, a conserved structural element in the Class A beta-lactamases, in which lies three other key residues identified by FunFams, near to the E166 - Ambler residues 157, 169 and 179, all differentially conserved in the 3 classes.

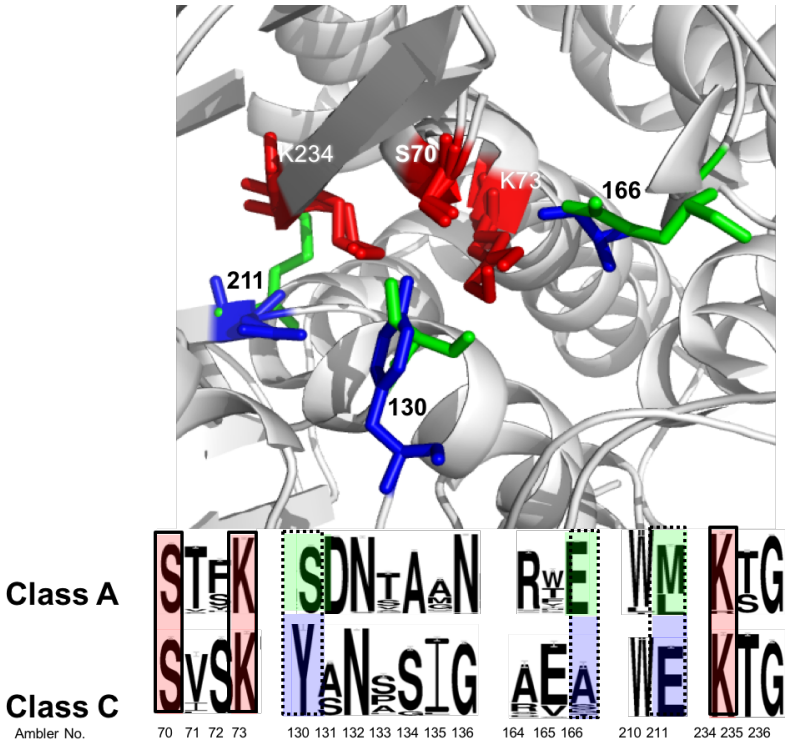


Figure 4.12: Functional determinants in Class A and Class C beta-lactamases. The three FDs - Ambler residue numbers 130, 166 and 211 are shown highlighted in Class A (green) and Class C (blue) beta-lactamase structures and sequence logos. The catalytic residues common to both enzymes (S70, K73 and K234) are shown in red.

Ambler residue 211

The Ambler residue 211, also a FunFam predicted functional determinant, is a highly conserved glutamate in Class C and usually a methionine residue in the other two classes. The E211 in Class C is located on the opposite side of the E166 in Class A beta-lactamases and is known to be involved in the hydrogen bonding network around the catalytic serine and affects the deacylation step to a small extent. Class A and Class C beta-lactamases are known to use opposite faces of the acyl-enzyme species for the approach of the hydrolytic water. The tyrosine at Ambler position 130 in Class C is implicated in activating water as mentioned above and it is likely that this tyrosine (lying in between E211 and S70, see Figure 4.12) assists the E211 in activating the water molecule. This

is necessary since E211 is rather distant from the catalytic S70 in the Class C beta-lactamases.

4.5.1.4 FunFams help identify known functional determinants

The validation of some of the FunFam predicted functional determinants (FDs) for serine beta-lactamases by experimental data reported in the literature demonstrates the power of the CATH FunFams to detect these sites and then exploit this information to correctly separate the three classes. Many of these residues appear to be involved in different strategies for activating the water molecule used for hydrolysis of the acylated beta-lactams. The other functional determinants that lie in close proximity of the catalytic residues (see Table 4.2) and have not yet been reported in the literature may be good targets for mutagenesis experiments to better characterise the reaction chemistry of the serine beta-lactamases.

4.6 Recent developments

Based on the validation of reasonable functional purity of the FunFams and the manual assessment of their utility to identify functional determinants in the serine beta-lactamases, we concluded that the FunFams provide useful sequence conservation information that could be used to help identify functional sites in proteins. This led to the recent development of a functional site prediction method, FunSite.

4.6.1 FunSite: identification of functional sites using FunFams

It has already been established in Chapter 2 that the CATH FunFams group together domain sequences that are functionally related and that the highly conserved residues in the FunFam multiple-sequence alignments (MSAs) are significantly enriched in functionally important residues (Das *et al.*, 2015b). Recent studies of selected FunFams by Garcia *et al.* (2016) have reported that the Fun-

Fams are also structurally coherent. Based on the sequence conservation information and structural data in the CATH FunFams, a method called FunSite was developed to identify functional sites in proteins that could also be used for improving the functional purity of the FunFams.

The FunSite method (see Figure 4.13) predicts functional sites in query protein sequences using sequence and structure information from the CATH FunFam MSAs, and is based on the assumption that functionally important residues tend to spatially cluster together in the three dimensional structure. It requires a protein domain sequence in FASTA format as input and a representative structure for the FunFam it matches best (i.e. lowest E-value match). The method provides a prediction of putative functional residues in the query domain as output. The FunSite method uses two separate protocols for the prediction of active site residues and ligand-binding residues.

In the first step of the FunSite method, the query sequence is scanned against the library of CATH FunFam HMMs using HMMER3 (Eddy, 2009). The query sequence is assigned to a FunFam if the E-value of the match to the FunFam HMM achieves the inclusion threshold of the FunFam (see Section 3.3.1 in Chapter 3). Subsequently, the method scores the conservation of residues in the matched FunFam alignment using Scorecons and maps the top ranked residues, according to the Scorecons score, to the structural representative of the FunFam. The top ranked residues are defined as residues ranked in the top N^{th} percentile where N can be 20 or 10 or can be specified by the user. The structural representative for a FunFam is chosen as the structural domain having the highest structural similarity to all structural domains within the FunFam.

4.6.1.1 FunSite protocol for prediction of active sites

Clefts are identified in the FunFam structural representative using Speedfill, a modified version of SURFNET (Laskowski, 1995), and only those top ranked

Figure 4.13: Overview of the FunSite method.

residues that have a Scorecons score of $\geq$ 0.8 and lie within a distance of 4 Å of the 5 largest clefts in the structural representative are selected. These selected residues are then clustered at a radius of 4 Å using a using a recursive clustering algorithm developed by Dr. Andrew Martin, Structural and Molecular Biology Dept., UCL, (personal communication) that clusters residues based on the minimum distance between all atoms of the residues in a protein structure. In recursive partitioning, the algorithm starts with a single cluster of residues and then splits it into multiple homogeneous clusters that have the smallest within cluster residue distances where the 'within cluster residue distance' is used as a measure of how homogeneous the cluster is with respect to the residues in it. Only clusters with at least 3 residues are retained. The number of residues in these clusters that have Scorecons scores $\geq$ 0.95 are then counted. Bartlett *et al.* (2002a) had reported that on average an enzyme has 3.5 catalytic residues and on manual inspection of the active sites of enzymes with known catalytic residue information, it was seen that the active site of an enzyme does not generally exceed 10 residues. As a result, if more than 10 FunSite predicted residues have Scorecons scores $\geq$ 0.95, only the top 10 FunSite residues are predicted as active site residues. Otherwise all residues within the clusters (with at least 3 residues) are predicted as active site residues.

4.6.1.2 FunSite protocol for prediction of ligand-binding sites

Solvent accessibility calculations for all residues are performed on the FunFam structural representative using NACCESS (Hubbard and Thornton, 1993) that is based on the program ACCESS (Richmond and Richards, 1978). Residues with relative accessible surface area (RSA) $>$ 20% are considered as solvent exposed residues and those with RSA $\leq$ 20% are considered as buried (Chen and Zhou, 2005; Chen *et al.*, 2013). Subsequently, the depth of each residue is calculated as the mean distance of all atoms of the residue from the nearest atom in the surface

of the structural representative. Residues with depths $\leq$ 5Å are considered as close to the protein surface and others are considered to be buried deep inside the structure (Tina *et al.*, 2007). Clefts are also identified in the FunFam structural representative using Speedfill (Laskowski, 1995).

The top ranked residues that lie close to the surface and any residue with a Scorecons score of $\geq$ 0.7 that lies within a distance of 1 Å of the 3 largest clefts in the structural representative are selected. The selected residues are then clustered at a radius of 5 Å using the recursive clustering algorithm developed by Dr. Andrew Martin, Structural and Molecular Biology Dept., UCL (personal communication). As described previously, this clusters residues based on the minimum distance between all atoms of the residues in a protein structure. Only clusters with at least 3 residues are retained and their constituent residues are predicted as ligand-binding residues.

The filtering of the top ranked residues within clefts or near the surface in the above FunSite protocols followed by spatial clustering aids the subdivision of large conserved surface patches into smaller and more distinct functional sites. The active site protocol, however, does not limit its analysis to surface residues. This is because a large number of known catalytic residues in experimental protein structures have been reported to have low relative solvent accessibilities ($<$ 7% relative surface accessibility) including a few catalytic residues that are completely buried (Bartlett *et al.*, 2002b).

4.6.1.3 Assessment of FunSite predictions

For assessing the performance of functional site predictions by FunSite, a dataset of 938 protein domain sequences in CATH (v4.0) was generated that had known structures and known catalytic site residues in the Catalytic Site Atlas (Porter *et al.*, 2004, version 2.0) identified from literature-based evidence. Biologically relevant ligand-binding residues for the dataset were extracted from the NCBI In-

ferred Biomolecular Interaction Server (IBIS) (Shoemaker *et al.*, 2012) excluding any residues inferred by homology.

HMM models were generated for all CATH FunFam MSAs. Query sequences had been excluded from these MSAs. The query domains were then scanned against the FunFam HMM library using HMMER3. 816 domains could be assigned to 371 FunFams of high information content i.e. FunFams with a DOPS score of greater than 70 (see Section 1.2.1.2 in Chapter 1, for explanation of DOPS score and threshold) from 225 superfamilies. The query set was then reduced to a representative set of 371 domains, containing only one domain for each unique CATH FunFam.

The FunSite protocol was run on the query domain sequences and a prediction of active site residues and ligand-binding residues were obtained for each query. The predicted functional site residues in the FunFam structural representative were then mapped to the known query domain structure by aligning the query domain sequence to the FunFam MSA containing the structural representative using the *mafft –add* option in MAFFT (Katoh *et al.*, 2002). To assess the performance of FunSite, none of the query domain structures were used as structural representatives by the protocol.

4.6.1.4 Residue enrichment analysis

Residue enrichment analyses, similar to those described in Chapter 2, Section 2.3.7, were performed on the query dataset to compare the enrichment of known catalytic and ligand-binding residues within the predicted FunSite residues compared to the background set of all residues in the query domains. For each query domain, enrichment scores were calculated as the difference between the proportion of predicted residues that are known functional residues (i.e. catalytic or ligand-binding) and the proportion of all residues that are known functional residues. The enrichment scores of query domains were averaged for each su-

perfamily for both catalytic and ligand-binding residues. Different top percentile ranked residues (20^{th} or 10^{th}) were used as definitions for top ranked residues in the FunSite protocol to evaluate their comparative performance in the query set. Enrichment analysis was also done for a FunFam protocol using Scorecons scores $\geq$ 0.7 (FunFams$_{Scorecons-0.7}$) to select functional residues as used in Section 2.3.7 in Chapter 2.

Enrichment of catalytic residues

Figure 4.14 shows the distribution of averaged enrichment scores for each superfamily, for 222 superfamilies represented by the query domains, for predicted FunSite catalytic residues. Higher enrichment scores for superfamilies indicate a higher enrichment of catalytic residues within the predicted FunSite residues

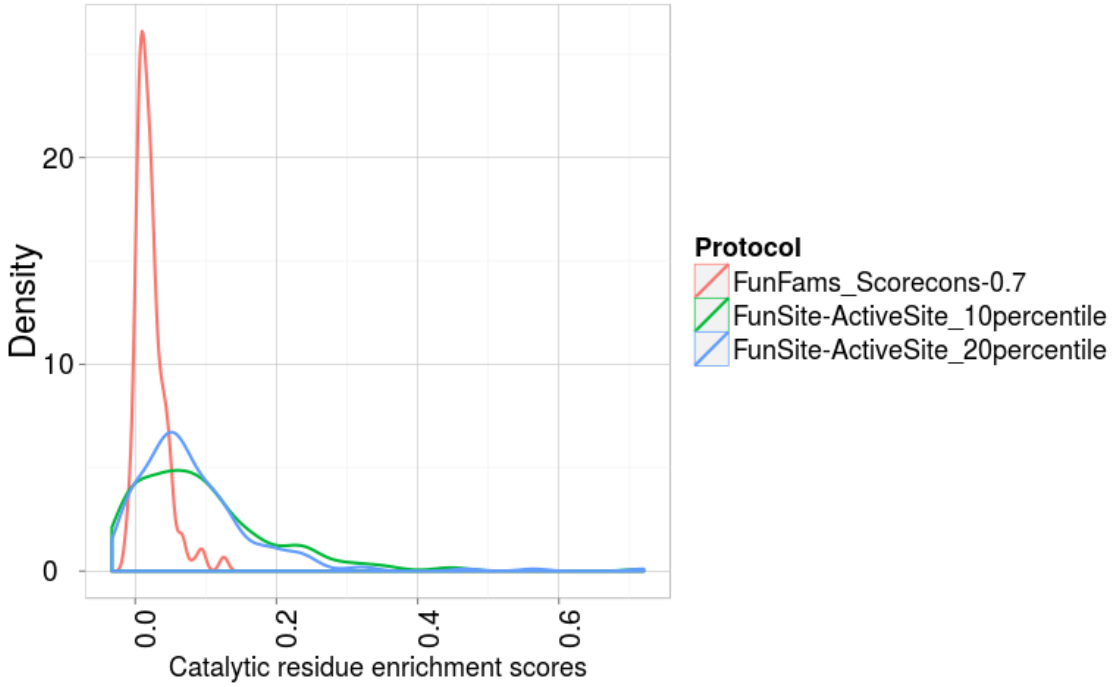


Figure 4.14: Distribution of averaged enrichment scores for each superfamily for predicted FunSite catalytic residues using 20^{th} and 10^{th} percentile ranked residues and residues predicted from FunFams$_{Scorecons-0.7}$. Using Wilcoxon Rank-Sum tests, both the two FunSite distributions were found to be significantly different from the FunFams$_{Scorecons-0.7}$ distributions with $p < 2.2\text{x}\,10^{-16}$. No significant difference was found between the FunSite 20^{th} and 10^{th} percentile protocol distributions.

in the query domains that have been assigned to the superfamily. Two-sided Wilcoxon Rank-Sum tests were used to determine whether there was a significant difference between the means in the distribution of average enrichment scores in the FunSite protocols using 20^{th} and 10^{th} percentile ranked residues and the FunFams$_{Scorecons-0.7}$ protocol. It was seen that overall, the FunSite method (using either 20^{th} or 10^{th} percentile ranked residues) provides a clear advantage ($p < 2.2\text{x}\,10^{-16}$) over simply selecting conserved sites (residues having Scorecons $\geq$ 0.7) from the FunFam MSAs for the prediction of active site residues.

Whilst, for more than 95% of the query domains in the dataset, the FunSite predicted active site residues (using both 20^{th} and 10^{th} percentile ranked residues) show higher enrichment for catalytic residues from CSA compared to the FunFam protocol of selecting residues using a Scorecons score cut-off of 0.7, for $\sim$ 2% of the query domains, the FunSite active site predictions were affected by the constraint of using spatial clusters containing 3 or more residues and for another 2% of the query domains, where predicted residues by FunSite or the FunFam protocol showed no enrichment of catalytic residues, the FunFam to which the query domain was assigned, was found to be functionally impure i.e. containing functionally different sequences.

An unpaired, one-sided Wilcoxon rank sum test (Kruskal, 1957) was also run on the averaged enrichment values for catalytic residues for all superfamilies using the wilcox.test function in R (R-Core-Team, 2014). This test assessed a *p*-value for the null hypothesis that the median enrichment value was zero. An enrichment value of zero indicates that the proportion of functional residues within predicted FunSite residues is the same as the proportion of functional residues within residues that are not predicted by FunSite. The alternative hypothesis assumed that the median enrichment value was greater than zero, i.e. a positive enrichment value, which reflected a greater proportion of functional residues within predicted FunSite residues in comparison to residues not predicted by FunSite.

Wilcoxon Rank-Sum tests reported significant p-values for FunSite predicted catalytic residues (Table 4.3).

Table 4.3: *p* values calculated from enrichment scores of catalytic residues (from CSA) within the predicted FunSite Active site residues and residues predicted from FunFams using Scorecons scores $\geq$ 0.7 compared to the background set of all residues in the query proteins using Wilcoxon Rank-Sum tests.

Protocol	**Enrichment of CSA residues**
FunSite using Top 20^{th} percentile residues	$p = 7.8 \text{ x } 10^{-31}$
FunSite using Top 10^{th} percentile residues	$p = 1.53 \text{ x } 10^{-28}$
FunFam residues with Scorecons score $\geq$ 0.7	$p = 1.5 \text{ x } 10^{-30}$

Enrichment of ligand-binding residues

Figure 4.15 shows the distribution of averaged enrichment scores for each superfamily, for 222 superfamilies represented by the query domains, for predicted FunSite ligand-binding residues. Two-sided Wilcoxon Rank-Sum tests were used to determine whether there was a significant difference between the means in the distribution of average enrichment scores in the FunSite protocols using 20^{th} and 10^{th} percentile ranked residues and the FunFams$_{Scorecons-0.7}$ protocol. It was seen that the FunSite method (using either 20^{th} or 10^{th} percentile ranked residues) provides a clear advantage ($p < 2\text{x } 10^{-8}$) over simply selecting conserved sites (residues having Scorecons $\geq$ 0.7) from the FunFam MSAs for prediction of ligand-binding site residues. Furthermore, the FunSite predictions using 10^{th} percentile ranked residues were found to provide better enrichment of ligand-binding residues than using the 20^{th} percentile ranked residues.

An unpaired, one-sided Wilcoxon rank sum test (Kruskal, 1957) was run on the averaged enrichment values for ligand-binding residues for all superfamilies using the wilcox.test function in R (R-Core-Team, 2014) similar to that was run for catalytic residues. Wilcoxon Rank-Sum tests reported significant p-values for ligand-binding residues predicted by FunSite (Table 4.4).

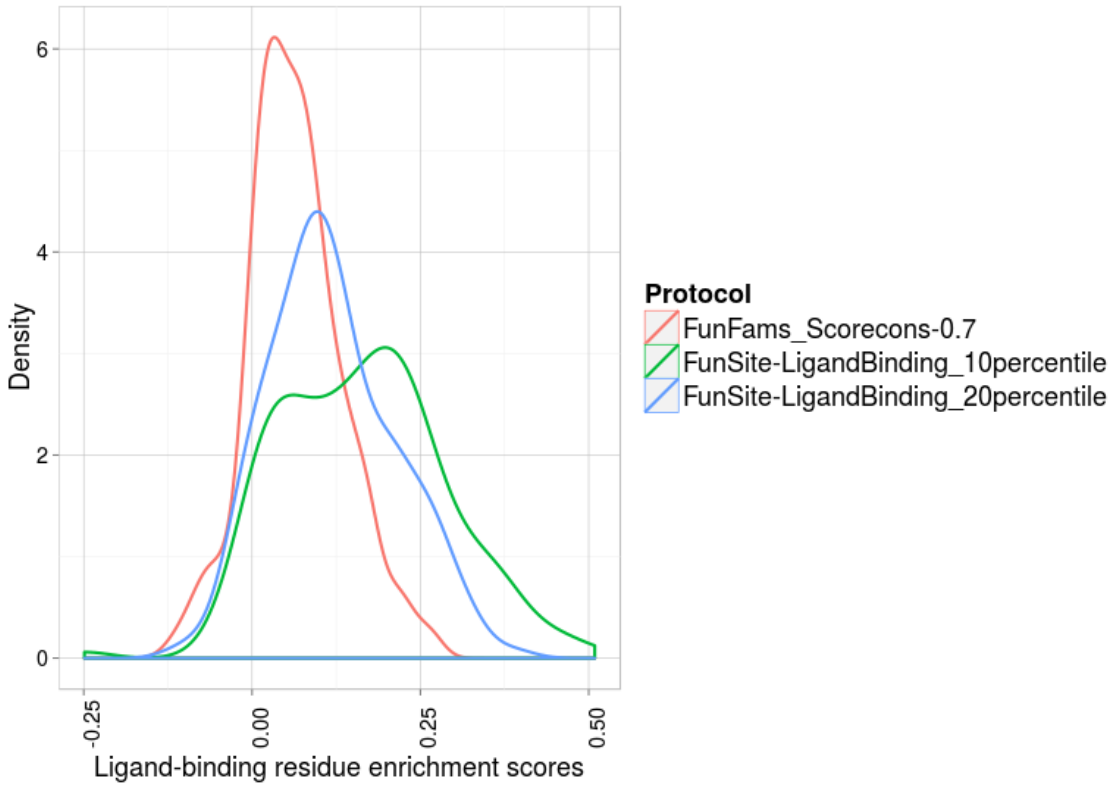


Figure 4.15: Distribution of averaged enrichment scores for each superfamily for predicted FunSite ligand-binding residues using 20^{th} and 10^{th} percentile ranked residues and residues predicted from FunFams$_{Scorecons-0.7}$. Using Wilcoxon Rank-Sum tests, both the two FunSite distributions and the FunFams$_{Scorecons-0.7}$ distributions were found to be significantly different from each other with $p < 2\text{x}\,10^{-8}$.

Table 4.4: *p* values calculated from enrichment scores of IBIS ligand-binding (LIG) residues within the predicted FunSite ligand-binding residues and residues predicted from FunFams using Scorecons scores $\geq$ 0.7 compared to the background set of all residues in the query proteins using Wilcoxon Rank-Sum tests.

Protocol	Enrichment of IBIS LIG residues
FunSite using Top 20^{th} percentile residues	$p = 2\,\text{x}\,10^{-27}$
FunSite using Top 10^{th} percentile residues	$p = 4.5\,\text{x}\,10^{-28}$
FunFam residues with Scorecons score $\geq$ 0.7	$p = 7\,\text{x}\,10^{-22}$

4.6.1.5 Comparison with Evolutionary Trace using MCC and BDT scores

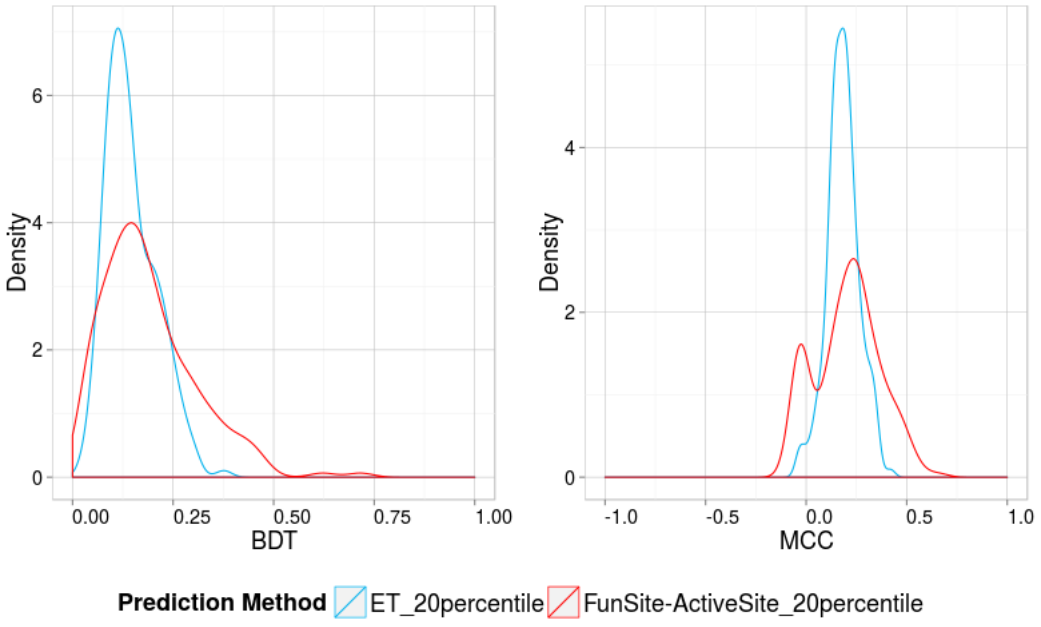
The performance of FunSite in predicting functional (catalytic and ligand-binding) residues was compared with the Evolutionary Trace (ET) method using MCC and BDT scoring measures. This was done by comparing the predicted FunSite residues for a query dataset using different top percentile (20^{th} and 10^{th} per-

centile) ranked residues against the corresponding top percentile ranked Evolutionary Trace residues. The pre-calculated Evolutionary Trace predictions for the query dataset was downloaded from the ET server (Lichtarge *et al.*, 1996; Lua *et al.*, 2016). The MCC and BDT scores (using a distance threshold of 3 Å), routinely used for assessment of functional site predictions (Schmidt *et al.*, 2011; Gallo Cassarino *et al.*, 2014), were calculated with respect to known catalytic residues in CSA and ligand-binding residues in IBIS. The comparisons were done for a common dataset of 246 query domains.

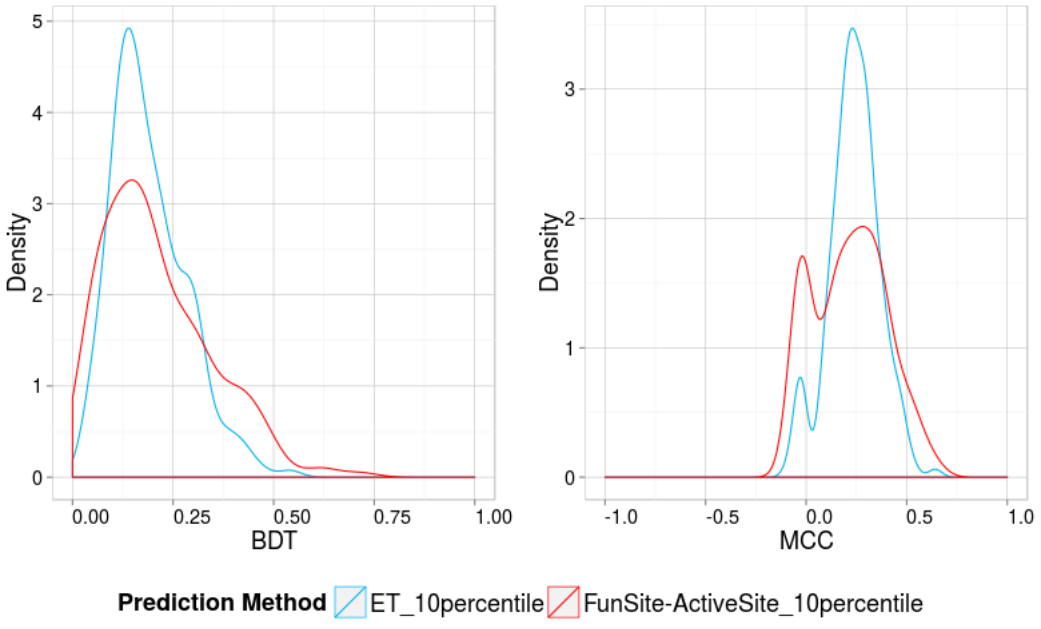
Figures 4.16 and 4.17 show the comparison of BDT and MCC score distributions for active site and ligand-binding site predictions respectively by FunSite and ET for a common dataset of 246 query domains with respect to known catalytic residues in CSA and ligand-binding residues in IBIS respectively. Two-sided Wilcoxon Rank-Sum tests were used to determine whether there was a significant difference between the means in the distributions of BDT or MCC scores for the predictions by ET and FunSite.

For active site site predictions, the FunSite predictions for the query domains was found to give higher BDT and MCC scores for a larger number of query domains compared to ET using both 20^{th} and 10^{th} percentile ranked residues (Figure 4.16). However, the difference between the means in the distributions of ET and FunSite predictions was significant only for the BDT score distributions of ET and FunSite using the top 20^{th} percentile ranked residues with a p value $<$ 0.000176 (Figure 4.16a).

For ligand-binding site predictions, the FunSite predictions for the query domains was found to give higher BDT scores for a larger number of query domains compared to ET using the top 10^{th} percentile ranked residues (Figure 4.17). The difference between the means in the distributions of ET and FunSite predictions using the top 10^{th} percentile ranked residues was found to be significant with a *p* value $<$ 0.0056 (Figure 4.17a). No significant differences was found in the other distributions of ET and FunSite for the ligand-binding site predictions.

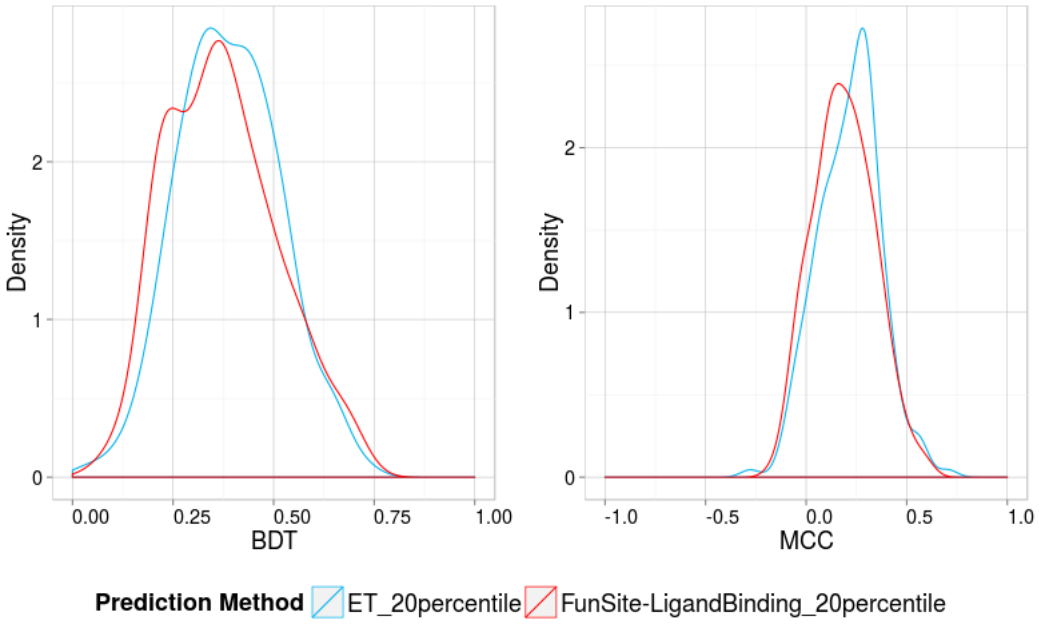


(a)

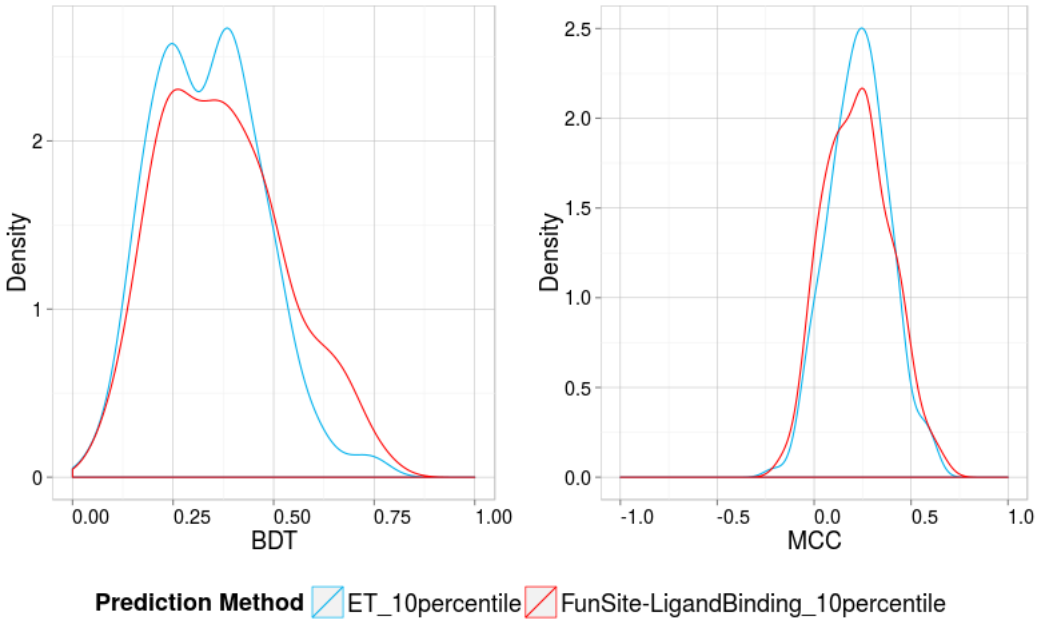


(b)

Figure 4.16: Distribution of BDT and MCC scores for predicted active site residues by FunSite and ET method using the top **(a)** 20^{th} and **(b)** 10^{th} percentile ranked residues.



(a)



(b)

Figure 4.17: Distribution of BDT and MCC scores for predicted ligand-binding site residues by FunSite and ET method using the top **(a)** 20^{th} and **(b)** 10^{th} percentile ranked residues.

4.6.2 Conclusion and future work

In this chapter an in-depth analysis of the functional families or FunFams in the CATH-Gene3D resource was first performed. The FunFams were then examined for their utility in exploring superfamily diversity and identifying functional sites in proteins.

The CATH(v4.0)-Gene3D(v12) resource classifies over 235,000 structural domains and over 25 million sequence domains into 2735 superfamilies. The superfamilies in CATH-Gene3D were sub-classified into functional families or FunFams using the GeMMA (Lee *et al.*, 2010) clustering algorithm and the FunFHMMer (Das *et al.*, 2015b, see Chapter 2) functional classification algorithm. Approximately 4.56 million Gene3D sequences that had at least one sequence relative annotated with high-quality GO annotations (see Section 2.3.2 in Chapter 2) at 90% sequence identity were used as seed sequences for the FunFHMMer protocol. This resulted in generation of 110,439 FunFams in 2735 superfamilies that could be mapped to over 11.7 million (45.73%) Gene3D sequences and 179,826 (76.52%) CATH structural domains.

The fact that less than half the Gene3D sequences are assigned to the FunFams could be attributed to the inability of seed sequences to represent the sequence diversity of some superfamilies and conservative inclusion thresholds of the FunFam models. Furthermore, although $\sim$ 82% of all the FunFam sequences could be mapped to highly informative FunFams (i.e. presence of evolutionary distant relatives in a FunFam), only about 16% of the FunFams were found to have high information content (i.e. as measured by the DOPS score calculated by Scorecons, see Section 1.2.1.2 in Chapter 1). Use of electronically annotated sequences as seed sequences and lowering the value of FunFam inclusion thresholds would help in increasing the information content and coverage (i.e. increase the percentage of Gene3D sequences assigned to a FunFam) of a large number of FunFams. In the current version of FunFHMMer, only sequence clus-

ters with at least one high-quality GO annotation are used along with conservative FunFam inclusion thresholds as the FunFams are primarily used for function prediction to ensure their functional purity. However, it may be useful to improve the information content and coverage of the FunFams for other applications such as prediction of functional sites and structural modelling.

The FunFams were used for exploring superfamily diversity in the CATH-Gene3D resource and it was seen that the functional classification of superfamilies into FunFams sheds light on superfamily diversities in terms of structure, function (measured by EC number diversity) and MDA. The top 200 superfamilies in CATH, ranked by the highest number of FunFams, were found to dominate nature, accounting for more than 65% of all CATH structural domains. An analysis of these top 200 superfamilies revealed that they show significantly higher diversity in structure, EC numbers and MDAs than the rest of the superfamilies. Furthermore, the enzyme-associated domain superfamilies among the top 200 superfamilies that contain domains that are solely responsible for enzyme catalysis were found to show significantly more diversity in structure, EC numbers and MDAs than the remaining top 200 superfamilies. Furthermore, visualisation of FunFam relationships in a superfamily using networks provided a good approach to examine and gain useful insights about the superfamily diversity.

The utility of the sequence conservation information of FunFams in identifying protein functional sites was manually assessed on serine beta-lactamases (enzymes that degrade beta-lactam antibiotics resulting in antibiotic resistance) that are assigned to the beta-lactamase/DD-peptidase superfamily that also contains DD-peptidases which are closely related to the serine beta-lactamases. The FunFams in the beta-lactamase/DD-peptidase superfamily were found to separate the three Classes of beta-lactamases (Classes A,C and D) and DD-peptidases into separate FunFams. Serine beta-lactamase functional determinants (i.e. residues differing between the three serine beta-lactamase Classes)

were identified by analysing a three-way structure based-sequence alignment of the three serine-lactamase Class FunFams using GroupSim (Capra and Singh, 2008). Detailed analysis of the functional determinants revealed that these residue positions are likely to contribute to the differences in the implementation of the catalytic mechanism of the three Classes of beta-lactamases.

Based on the functional purity and predictive power of the FunFams, the FunSite method was developed that makes use of the sequence conservation information in FunFams and associated structural data to predict functional sites in protein domains. Using a dataset of 371 query domains, the FunSite predicted active site residues and ligand-binding residues were found to be significantly enriched in catalytic residues from the CSA and IBIS ligand-binding residues. A comparison of the FunSite and Evolutionary Trace (ET) predictions on a common dataset of 246 protein domains showed that the FunSite method performs competitively with the widely-used ET method. The performance of the FunSite method may be further improved by using a more sophisticated method for ranking residues that does not allow repetitive ranks. Furthermore, the relative ranking of residues can be improved by differentiating between conserved residues that change between FunFams (i.e. functional determinants) and conserved residues that are invariant in related FunFams or the entire superfamily. Furthermore, it was found that the ligand-binding residues predicted by the FunSite method showed a significant enrichment of protein-protein interaction (PPI) residues with a *p* value $< 2.14\text{x}10^{-6}$ compared to residues with Scorecons score $\geq$ 0.7 in FunFams (p = 0.143) using a Wilcoxon Rank-sum test. This suggests that it may be possible to extend the FunSite ligand-binding site prediction method in future to predict PPI residues by removing the constraints of pockets and focusing on surface patches.

Chapter 5

Structure-based classification and annotation of moonlighting proteins

5.1 Background

An increasing number of proteins have been identified to be multi-functional (performing multiple functions) in the post-genomic era contrary to the over simplistic one gene-one enzyme-one function hypothesis prevalent in the 1940s (Beadle and Tatum, 1941). A number of these proteins have been found to moonlight, i.e., perform multiple independent functions within a single polypeptide chain (Figure 5.1) that are not due to gene fusions, splice variants or post-translational modifications (Jeffery, 1999). The multiple roles of moonlighting proteins are not restricted to certain organisms or protein families, nor do they have a common mechanism through which they switch between different functions. Orthologous proteins in different organisms do not necessarily share moonlighting functions.

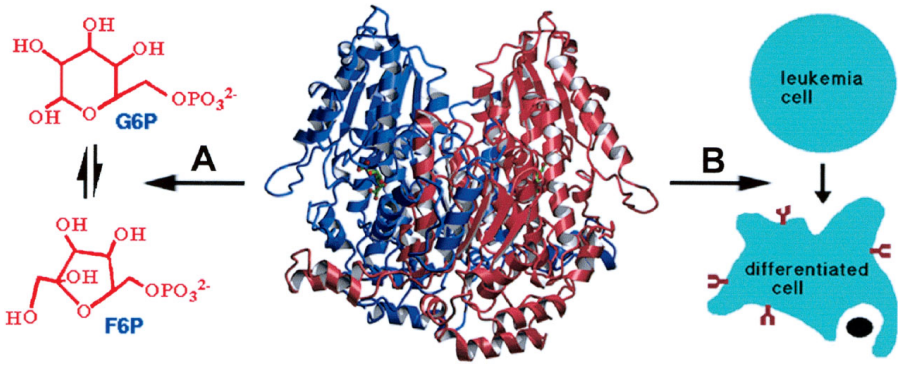


Figure 5.1: Example of a moonlighting protein. The figure shows the structure of the rabbit Phosophoglucose isomerase along with two arrows (A and B) indicating its primary and moonlighting functions respectively. The arrow A indicates its primary function as a cytosolic enzyme (catalysing the interconversion of glucose-6-phosphate and fructose-6-phosphate) and the arrow B indicates its moonlighting function as an extracellular cytokine that causes the differentiation of leukemia cells. Taken from Jeffery (2003).

Experimentally identified moonlighting proteins have been shown to switch functions as a consequence of changes in cellular locations within and outside the cell, expression in different cell types, oligomerisation states, ligand binding locations, binding partners and complex formation (Figure 5.2) (Jeffery, 1999, 2004b).

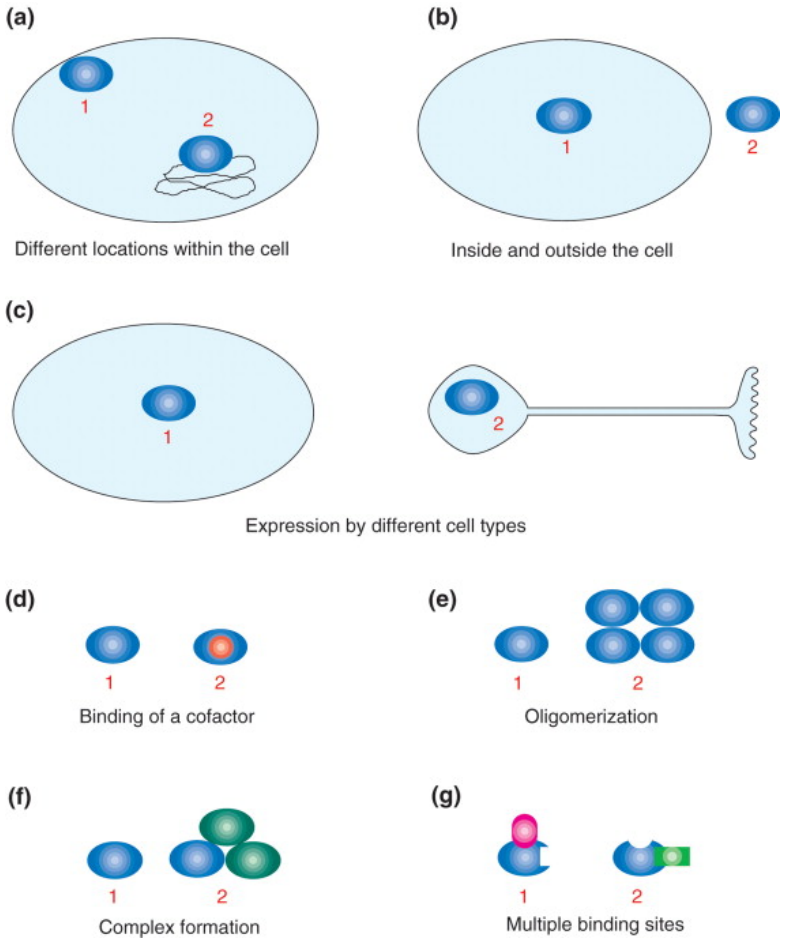


Figure 5.2: Examples of mechanisms in moonlighting protein for switching between the primary and moonlighting functions. Taken from Jeffery (2003).

A large number of moonlighting proteins have been found to be involved in bacterial virulence, DNA synthesis or repair, cancer cell motility and angiogenesis, amongst other processes. For example, Neuropilin is a moonlighting protein that is known to show diverse functions due to changes in cellular contexts. In endothelial cells, it is a vascular endothelial cell growth factor (VEGF) receptor

and a major regulator of angiogenesis, vasculogenesis and vascular permeability. However, in nerve axons, it is a receptor for a different ligand (Semaphorin III) and mediates neuronal cell guidance.

Currently, there exist two manually-curated databases of moonlighting proteins, MultitaskProtDB (Hernández *et al.*, 2014) and MoonProt (Mani *et al.*, 2014), each of which lists more than 280 moonlighting proteins known in the literature. However, the rapid increase in the number of identified moonlighting proteins suggest that the phenomenon may be common in all kingdoms of life. So far, the moonlighting function(s) of the proteins have been mostly discovered by serendipity and little is known about the molecular mechanisms of proteins (Jeffery, 2004a). Consequently, any efforts to characterize the molecular mechanisms of such proteins and understand their structure-function relationship would aid in identifying new moonlighting functions and help in better understanding of the complex functional role of proteins in the cell.

5.2 Identification of moonlighting by computational approaches

Moonlighting proteins pose a major challenge for automated protein function prediction methods as a majority of these methods use homology to inherit annotations. Homology-based function prediction methods can often lead to erroneous annotations of a query protein if a moonlighting function is inferred from a homologous protein. Similarly, annotations inferred by these methods for a query moonlighting protein would be incomplete if inherited from an homologous non-moonlighting proteins.

A number of computational tools have been assessed to understand whether current approaches for protein function prediction can identify the moonlighting functions of proteins (Gómez *et al.*, 2003; Khan *et al.*, 2012; Hernández *et al.*, 2015). These methods can be broadly categorised into: (i) remote homology

search using PSI-BLAST (Altschul *et al.*, 1997), (ii) motif or domain search-based methods using data from protein family resources such as BLOCKS, ProDom, Pfam and other InterPro member databases, (iii) structure-based methods and (iv) methods using interactomics data from protein-protein interaction (PPI) databases such as DIP (Xenarios *et al.*, 2002), BIND (Bader *et al.*, 2003) and APID (Prieto and De Las Rivas, 2006). While the first three categories use either sequence or structural homology to predict protein functions, the last category either exploits PPI data to search for interaction partners for a query protein to predict its functions (Espadaler *et al.*, 2008; Gómez *et al.*, 2011) or clusters PPI networks to identify novel multifunctional or moonlighting proteins (Becker *et al.*, 2012).

Gómez *et al.* (2003) and Hernández *et al.* (2015) compared the performance of the PSI-BLAST and various motif/domain search against protein family resources manually on a dataset of 46 and 288 moonlighting proteins respectively. Both studies concluded that remote homology searches using PSI-BLAST gives good performance for identifying moonlighting proteins, however, among the domain-based methods, Gómez *et al.* (2003) found ProDom to perform the best, while Hernández *et al.* (2015) found Pfam to perform best. Hernández *et al.* (2015) further concluded that PSI-BLAST results combined with information from PPI databases was found to give the best performance.

Khan *et al.* (2012) showed that the PFP (Protein Function Prediction) (Hawkins *et al.*, 2009) and ESG (Extended Similarity Group) (Chitale *et al.*, 2009) methods (only available as webservers) outperform PSI-BLAST in predicting diverse functions of moonlighting proteins using a small dataset of 19 moonlighting proteins using precision-recall curves similar to those used in CAFA (see Section 3.1.3.1 in Chapter 3). For the moonlighting protein dataset, Khan *et al.* (2012) had reported that ESG shows the highest precision in predicting GO terms while PFP provides higher coverage in predicting diverse GO terms associated with the proteins in the dataset. Both PFP and ESG can be regarded as modification of the

PSI-BLAST algorithm. The PFP method uses all the sequence hits of a PSI-BLAST search for a query sequence up to an E-value of 100 and combines the frequency of GO terms of all the sequence hits to predict GO terms using an E-value based scoring scheme along with a data-mining tool, Function Association Matrix, that predicts additional GO terms for the sequence hits from PSI-BLAST based on the frequency at which they co-occur in UniProt sequences (Hawkins *et al.*, 2006). In contrast, the ESG method performs iterative PSI-BLAST searches from the sequence matches of an initial PSI-BLAST search for a query sequence up to an E-value of 1000, performing a multi-level exploration of the sequence similarity space around the query sequence. It predicts GO terms by combining information from all the sequence matches in the sequence similarity space in the vicinity of the query sequence (i.e. up to an E-value of 100) using the data-mining tool similar to the PFP algorithm. While the PFP predictions are designed to provide a larger coverage by retrieving annotations widely from weakly similar sequences, ESG predictions provide higher specificity from consistently predicted GO terms in an iterative PSI-BLAST search (Khan *et al.*, 2015).

5.3 Aims and Objectives

This chapter first proposes a classification of moonlighting proteins based on the structure-function analyses of selected moonlighting proteins. A few examples of moonlighting proteins in each classification are described in detail followed by some general trends. Secondly, we assess the performance of the FunFHMMer function prediction pipeline for functional annotation of moonlighting proteins. This has been published in:

Das, S. and Orengo, C. A. (2015). Protein function annotation using protein domain family resources, *Methods*, 93, 24–34.

5.4 A structure-based classification of moonlighting proteins

A dataset of 23 known moonlighting proteins was constructed from the MoonProt (Mani *et al.*, 2014) database and the literature that had structural data and information regarding experimentally verified functional sites responsible for the primary and moonlighting function(s) of the protein was available. Information on catalytic site residues for the proteins were extracted from the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004) and additional functional annotations were extracted from PDBsum (de Beer *et al.*, 2014) and UniProtKB/Swiss-Prot (UniProt-Consortium, 2015).

The structural data available for the dataset of moonlighting proteins was examined and a classification of moonlighting proteins was proposed based on the spatial locations of the experimentally-verified functional sites exploited by a protein to perform its primary and moonlighting function(s). The primary and moonlighting function(s) of the proteins are referred to as 'Function 1' and 'Function 2' in this Chapter. The moonlighting proteins were classified in the following five categories:

i) Proteins having distinct sites for different functions in the same domain.

ii) Proteins having distinct sites for different functions in different domains.

iii) Proteins using the same residues for different functions.

iv) Proteins using different residues in the same or overlapping site for different functions.

v) Proteins using different structural conformations for different functions.

5.4.1 Proteins with distinct sites for different functions in the same domain

These are single domain or multi-domain proteins (listed in Table 5.1) that use distinct spatial functional sites of a single domain for carrying out their primary and moonlighting function(s).

5.4.1.1 α-Enolase (*Streptococcus pneumonia*)

α-Enolase (EC 4.2.1.11) from *Streptococcus pneumonia* is a key glycolytic enzyme (Function 1) that is also expressed on the bacterial cell-surface where it binds human plasminogen to facilitate the host invasion process during infection (Function 2) (Ehinger *et al.*, 2004). The protein is known to exist in an octameric state both in the cytoplasm and on the cell surface. Each monomer of α-enolase consists of a TIM barrel domain and a 2-layer $\alpha\beta$ sandwich domain (Figure 5.3).

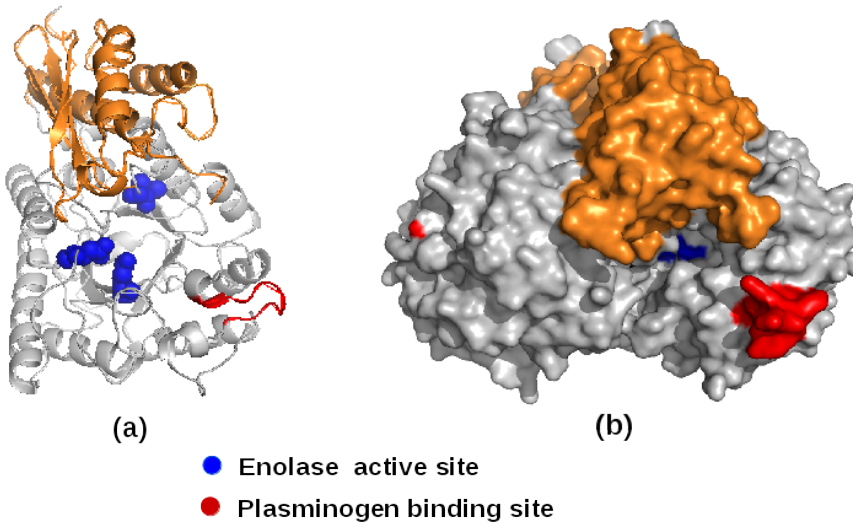


Figure 5.3: Primary and moonlighting functions of α-Enolase (PDB:1W6T). (a) Single chain of Enolase showing the enzyme active site (primary function) in blue and the plasminogen binding site (moonlighting function) in red. (c) Enolase monomer displayed as surface. The TIM barrel domain is coloured in grey (TIM barrel) and the 2-layer $\alpha\beta$ sandwich domain is coloured in orange.

Table 5.1: Proteins having distinct sites for different functions located in the same domain

Protein (Organism)	Function 1	Function 2	Structure	Refs.
Enolase (*S. pneumoniae*)	Enolase (EC 4.2.1.11)	Binds Plasminogen	1W6T	Ehinger *et al.* (2004)
Albaflavenone monooxygenase (*S. coelicolor A3(2)*)	Albaflavenone monooxygenase (EC 1.14.13.106)	Terpene synthase (EC 4.2.3.47)	3EL3	(Zhao *et al.*, 2008, 2009)
MAPK1/ERK2 (*H. sapiens*)	Mitogen-activated protein kinase 1 (EC 2.7.11.24)	Transcriptional repressor (binds DNA)	4G6N	(Hu *et al.*, 2009)
1-Cys Peroxiredoxin (*H. sapiens*)	Phospholipase A2 (EC 3.1.1.4)	Glutathione peroxidase (EC 1.11.1.15)	1PRX	(Fisher, 2011)
Cytochrome C (*S. cerevisiae*)	Electron carrier protein in Electron transport chain	Promotes Apoptosis (binds Apaf-1)	1YCC	(Lim *et al.*, 2002)
GCN4 (*S. cerevisiae*)	Transcription factor (binds DNA)	Ribonuclease (EC 3.1.27.5)	1YSA	Nikolaev *et al.* (2010)
I-AniI (*A. nidulans*)	Homing endonuclease (EC 3.1.-.-)	Transcriptional repressor (binds DNA)	3EH8	(Jeffery, 2004b; Takeuchi *et al.*, 2009)

The structurally conserved α-enolase active site is located in a surface pocket of the TIM barrel which comprises the catalytic residues Glu164, Glu205 and Lys342 in *S. pneumoniae*. Two plasminogen-binding sites have also been found in the TIM barrel domain at sites distinct from the active site which include a nine-residue internal motif ($_{248}FYDKERKYV_{256}$) and terminal lysine residues ($_{433}KK_{434}$). The former site has been shown to have a more important role in interacting with plasminogen than the latter. The last lysine residue is not part of the globular structure as it is disordered.

5.4.1.2 Albaflavenone monooxygenase, *(Streptomyces coelicolor A3(2))*

Albaflavenone monooxygenase (CYP170A1) (EC 1.14.13.106) from *Streptomyces coelicolor A3(2)*, catalyzes the last two steps in the biosynthesis of the antibiotic Albaflavenone (Function 1) (Zhao *et al.*, 2008). Study of the crystal structure of Albaflavenone monooxygenase has shown that it exists as a dimer having two chains, each consisting of a single, α orthogonal bundle (Figure 5.4). These studies also revealed that this protein can also function as a Terpene synthase (EC 4.2.3.47) (Zhao *et al.*, 2009) in the synthesis of farnesene isomers from farnesyl diphosphate (Function 2). This catalytic activity was identified on the basis of signature sequences and motifs associated with terpene synthases. The residues Trp92, Pro274, Val338, Ile447 and Thr448 are involved in the monooxygenase activity whereas the residues Arg116, Leu244, Leu248, Glu263, Val268, Leu271, Ile272 and Phe415 are associated with the terpene synthase activity and are located in different pockets in the protein. The monooxygenase activity was found to be optimal between the pH 7.2-8 and was found to decline at lower pHs which favours the terpene synthase activity (pH 5.5-6.5). This suggests that the two different enzymatic states of the protein possess optimal conformations at distinct pHs.

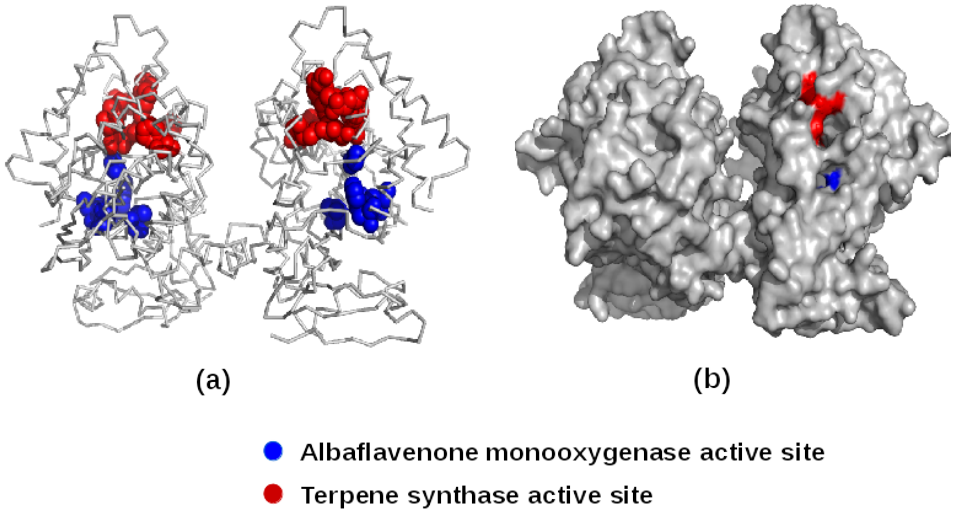


Figure 5.4: Primary and moonlighting functions of Albaflavenone monooxygenase (PDB: 3EL3). The monooxygenase (primary function) and terpene synthase (moonlighting function) active sites are shown in blue and red respectively in the (a) cartoon and (b) surface representation of Albaflavenone monooxygenase.

5.4.1.3 MAPK1/ERK2 (*Homo sapiens*)

Studies to characterise the human protein-DNA interactome revealed that the human Mitogen-activated protein kinase 1 (MAPK1) (also known as Extracellular signal-r kinase 2, ERK2) (Function 1) also functions as a DNA binding transcriptional repressor (Function 2) that regulates interferon gamma signalling (Hu *et al.*, 2009). The crystal structure of the human MAPK1 exists as a monomer which contains two domains: a discontinuous $\alpha\beta$ 2-Layer sandwich domain and a mainly α orthogonal bundle domain (Figure 5.5). The kinase active site residues involve Asp147, Lys149, Ser151 and Asn152. However, the motif involved in binding DNA is – ${}_{259}KARNYLLSLPHKNKVPWNR_{277}$ and it can be seen from Figure 5.5 that the kinase active site is located far from the DNA-binding motif.

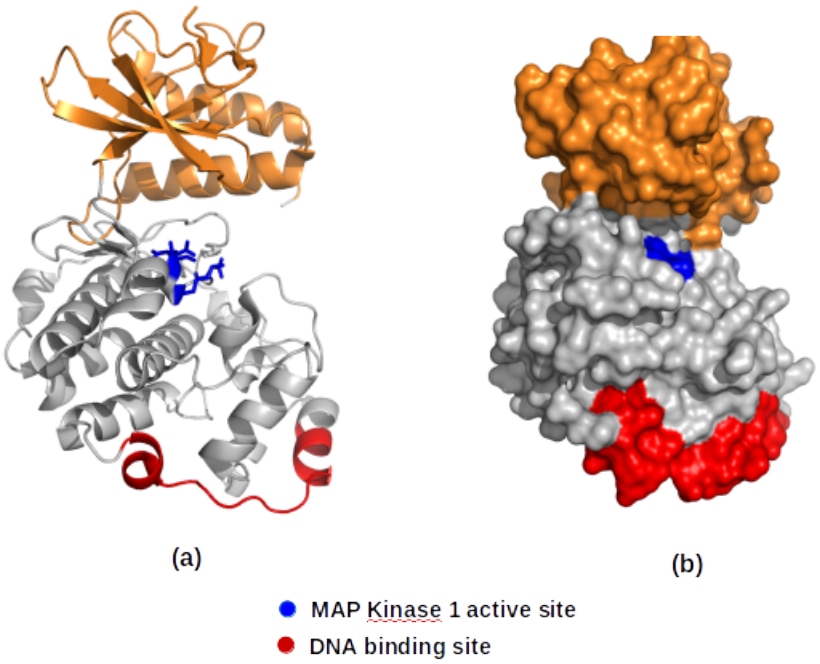


Figure 5.5: Primary and moonlighting functions of human MAPK1/ERK2 (PDB: 4G6N). The MAPK1 active site (primary function) is shown in blue and the DNA-binding motif (moonlighting function) is highlighted in red. Different domains are shown in grey and orange.

5.4.2 Proteins with distinct sites for different functions in different domains

The second category of moonlighting proteins are multi-domain proteins (listed in Table 5.2) which use functional sites in separate domains to carry out their primary and moonlighting function(s).

5.4.2.1 Malate synthase (*Mycobacterium tuberculosis*)

Malate synthase (EC 2.3.3.9) is a cytoplasmic enzyme (Function 1) involved in the glyoxalate pathway (Tolbert, 1981). In *M. tuberculosis* it has also been found on the cell wall, adapted to function as an adhesin that binds laminin and fibrinogen and this may contribute to *M. tuberculosis* virulence by promoting infection and dissemination (Function 2) (Anstrom and Remington, 2006). The structure of the *M. tuberculosis* malate synthase consists of two identical chains each of

Table 5.2: Proteins having distinct sites for different functions in different domains

Protein (Organism)	**Function 1**	**Function 2**	**Structure**	**Refs.**
Malate synthase *(M. tuberculosis)*	Malate synthase (EC 2.3.3.9)	Binds laminin	2GQ3	(Kinhikar *et al.*, 2006)
BirA (*E. coli*)	Biotin holoenzyme synthetase (EC 6.3.4.15)	Bio repressor	1BIB	(Wilson *et al.*, 1992)
MRDI (*H. sapiens*)	MTR-1-P isomerase (EC 5.3.1.23)	Mediator of cell invasion	4LDQ	(Templeton *et al.*, 2013)
Hexokinase 2 *(S. cerevisiae)*	**Hexokinase 2** EC 2.7.1.1)	Glucose sensor (interacts with transcriptional repressor Mig1)	1IG8	Gancedo and Flores (2008)
Neuropilin-I *(H. sapiens)*	Semaphorin binding	VEGF binding	2QQN	(Appleton *et al.*, 2007)
ATF2(*H. sapiens*)	Transcription factor	DNA damage response	1T2K	(Bhoumik *et al.*, 2005)

which consists of 4 domains - an α orthogonal bundle, a TIM barrel, a mainly β complex domain and an α up-down bundle (Anstrom and Remington, 2006). The malate synthase active site residues are Glu273, Asp274, Arg339, Glu434, Leu461, Asp462 and Glu633 (highlighted as blue sticks) and the residues that are associated with binding laminin or fibrinogen are Gln696-Glu727 (highlighted in red) (Figure 5.6). Both the sites are present in different domain regions of the protein.

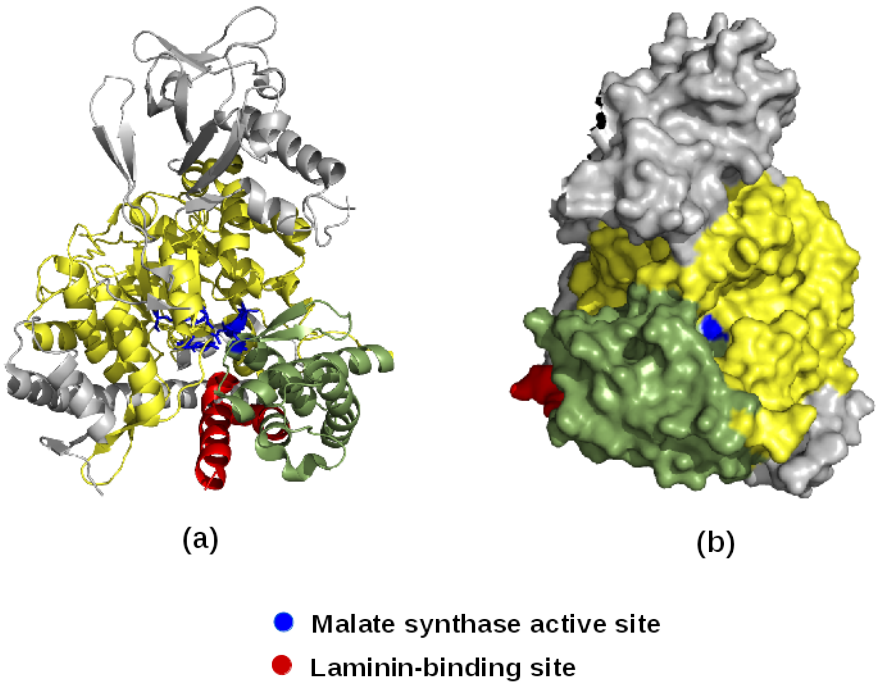


Figure 5.6: Primary and moonlighting functions of Malate Synthase (PDB: 2GQ3). The enzyme active site (primary function) is shown in blue and the laminin-binding site (moonlighting function) is shown in red. Different domains are shown in yellow, green and grey.

5.4.2.2 BirA (*E. coli*)

The *E. coli* BirA protein performs different functions depending on its dimeric state (Wilson *et al.*, 1992). As a heterodimer with biotin carboxyl carrier protein (BCCP) subunit of acetyl-CoA carboxylase, it functions as a biotin protein ligase (Function 1), and as a homodimer, it functions as a biotin operon repressor (Function 2) that binds to DNA (Jeffery, 2011). The BirA structure consists of three domains:

an α orthogonal bundle, an α/β 2-Layer sandwich domain and a mainly β SH3-type fold (Figure 5.7). The residues responsible for the two functions of BirA are located in distinct sites in the protein. The catalytic residues for the ligase activity of the protein are Arg118, Lys183 and Arg317 and are found in a pocket formed between the $\alpha\beta$ sandwich, the SH3 domain and a helix-turn-helix (H-T-H) motif (residues 22-46). This H-T-H motif, found in the orthogonal α bundle, is responsible for the DNA-binding function of the protein (Figure 5.7) (Wilson *et al.*, 1992).

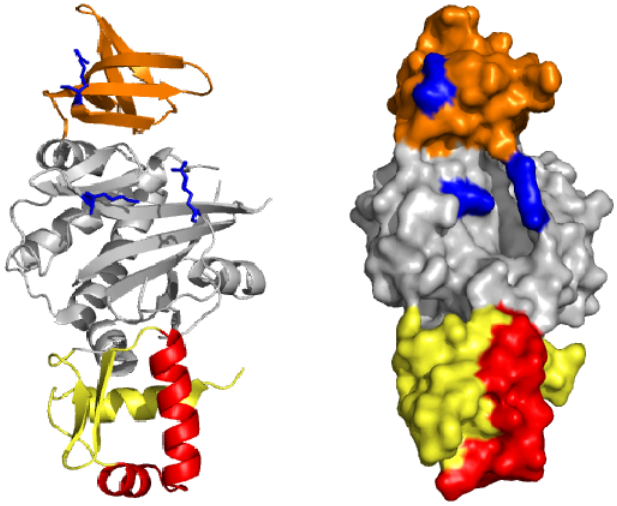


Figure 5.7: Primary and moonlighting functions of BirA (PDB: 1BIB). The enzyme catalytic site (primary function) residues are shown in blue while the H-T-H motif involved in binding DNA (moonlighting function) is shown in red. Different domains are shown in yellow, orange and grey.

5.4.2.3 MRDI (*H. sapiens*)

The protein, Mediator of RhoA-dependent invasion (MRDI), is a moonlighting protein found in humans (Templeton *et al.*, 2013) that acts as a methylthioribose-1-phosphate (MTR-1-P) isomerase (EC 5.3.1.23) (Function 1) and a mediator of melanoma cell invasion (Function 2) in melanoma cells. The MRDI structure consists of 2 chains each comprising a 4-helix bundle and a Rossmann fold (Figure

5.8). The catalytic residues of MRDI are Cys168 and Asp248 (shown in blue), which are located in the base of a pocket. Structural comparison of MRDI with other MTR-1-P isomerases and mutational analysis identified a site (comprising the residues Ser283 and Arg109 that are shown in red in Figure 5.8) responsible for the invasion phenotype that is distal from the catalytic site in another pocket of the protein formed between the two domains of each chain.

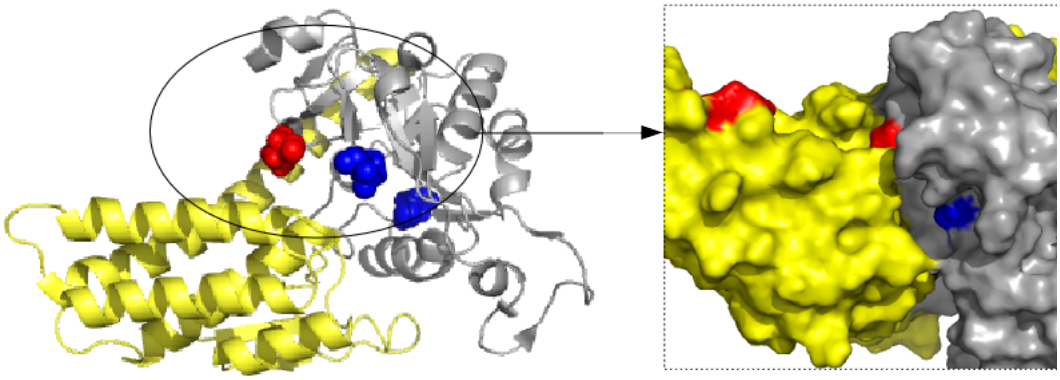


Figure 5.8: Primary and moonlighting functions of human MRDI (PDB:4LDQ). The active site residues (primary function) are shown in blue while the residues implicated in controlling invasion (moonlighting function) is shown in red. Different domains are shown in yellow and grey.

5.4.3 Proteins using the same residues for different functions

These are multi-domain proteins (listed in Table 5.3) which utilize the same functional site for carrying out their primary and moonlighting function(s).

Table 5.3: Proteins using the same residues for different functions

Protein (organism)	Function 1	Function 2	Structure	Refs.
GAPDH (*E. coli*)	GAPDH (EC 1.2.1.12)	NAD ribosylating activity	1DC5	(Aguilera *et al.*, 2010)
Leukotriene A-4 hydrolase *(H. sapiens)*	Leukotriene A-4 hydrolase (EC 3.3.2.6)	Aminopeptidase (EC 3.4.11.24)	2R59	(Haeggström, 2004)
Hemagglutinin (*Paramyxovirus*)	Hemagglutinin binding	Neuraminidase (EC 3.2.1.18)	1E8T	(Crennell *et al.*, 2000)

5.4.3.1 GAPDH (*E. coli*)

The *E. coli* glyceraldehyde-3-phosphate dehydrogenase (GAPDH; EC 1.2.1.12) (Function 1) is a multifunctional housekeeping protein. It also catalyses its own NAD^+-dependent ADP-ribosylation which has been implicated in host-pathogen interactions (Function 2) (Aguilera *et al.*, 2010). GAPDH consists of two chains, each comprising a Rossmann fold and a $\alpha_3\beta_5$ sandwich domain (Yun *et al.*, 2000). The three catalytic residues of GAPDH are Cys149, His179 and Ser238, which are located in the sandwich domain (Figure 5.9). However, mutational analyses have shown that the catalytic Cys149 (shown in red) is also the target residue of the ADP-ribosylation .

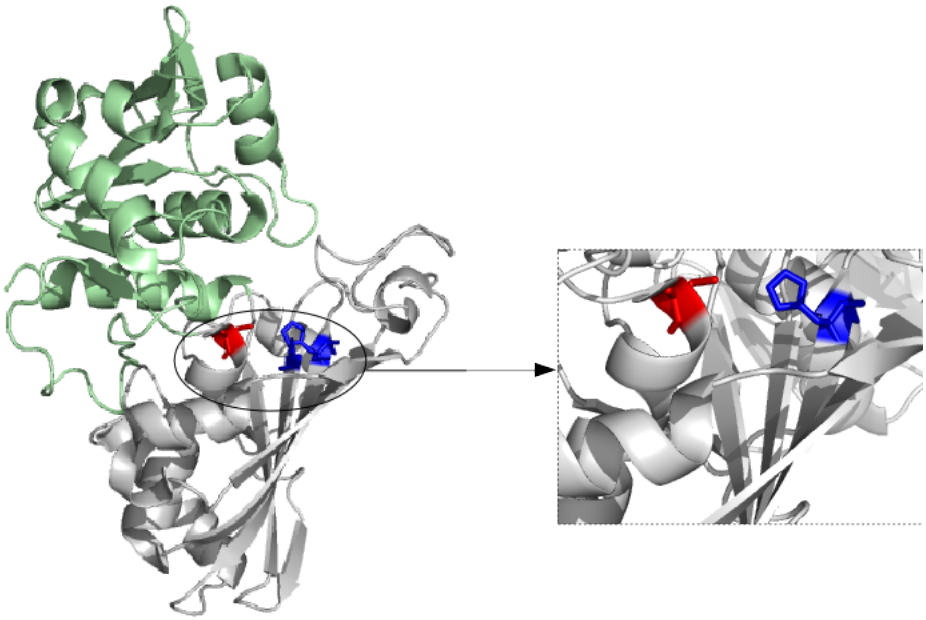


Figure 5.9: Primary and moonlighting functions of GAPDH (PDB: 1DC5). The catalytic site residue Cys149 (shown in red) is the residue known to be involved for both the primary and moonlighting functions of *E. coli* GAPDH. The other catalytic residue His179 is shown in blue.

5.4.3.2 Leukotriene A4 hydrolase (*Homo sapiens*)

Leukotriene A4 hydrolase (EC 3.3.2.6) is a bifunctional zinc metalloenzyme that converts the fatty acid epoxide leukotriene A4 (LTA4) into a potent chemoattractant, Leukotriene B4 (LTB4) (Function 1) and also exhibits an anion-dependent aminopeptidase activity (EC 3.4.11.24) (Function 2) (Haeggström, 2004). Both the enzymatic activities require the presence of the catalytic zinc which is coordinated by the three zinc-binding residues His295, His299, and Glu318. The crystal structure of the LTA4 hydrolase consists of 3 domains: a β sandwich, an α orthogonal bundle and a α-α superhelix domain (Figure 5.10) (Tholander *et al.*, 2008). It also contains the ${}_{269}GXMEN_{272}$ motif in the M1 family of zinc peptidases. Mutation of either of the catalytic residues Glu296 and Tyr383 residues resulted in loss of the aminopeptidase activity and mutation of the catalytic residue Glu271 abolished both the epoxide hydrolase activity and the aminopeptidase activity. Glu271 is a unique example of a catalytic residue that has distinct roles in two separate catalytic reactions for two chemically different substrates. Based on the LTA4 hydrolase structure and structure activity studies, mechanistic models for the role of Glu271 in the epoxide hydrolase activity and in the aminopeptidase reaction (Rudberg *et al.*, 2002) were proposed. In the first model for the epoxide hydrolase activity, Glu-271 activates a water molecule bound to the zinc to promote an acid-induced opening of the epoxide and the generation of a carbocation intermediate. In the second model for the aminopeptidase activity, Glu-296 polarizes a water molecule for nucleophilic attack at the carbonyl carbon of the scissile peptide bond. Tyr-383 donates a proton to the peptide nitrogen, and Glu-271 binds the terminal amino group.

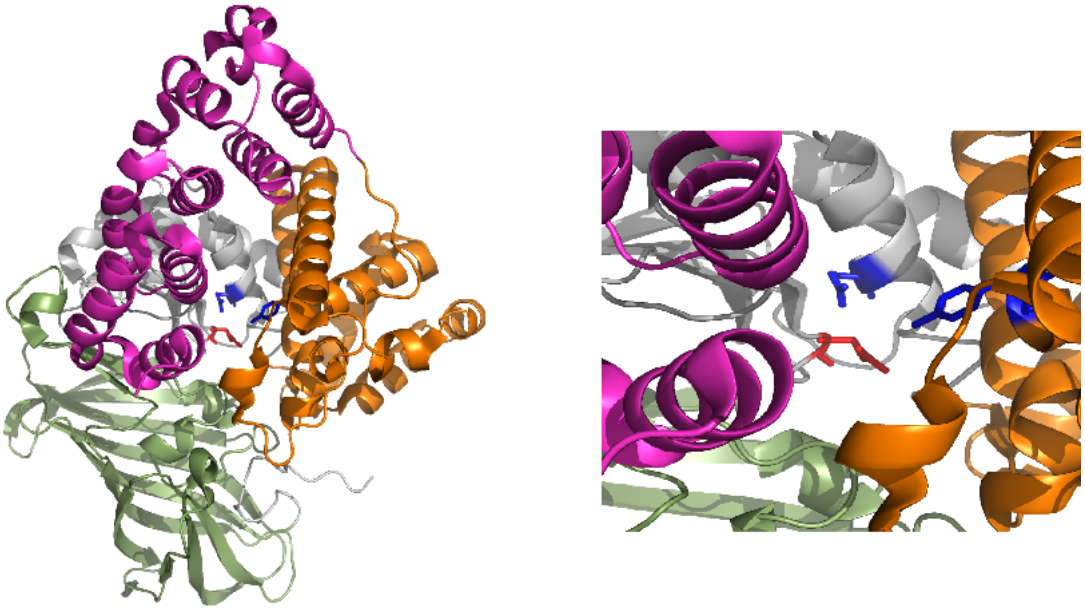


Figure 5.10: Primary and moonlighting functions of Leukotriene A4 Hydrolase (PDB: 2R59). The catalytic site residue Glu271 that is involved in two different catalytic reactions - epoxide hydrolase (primary function) and amino-peptidase (moonlighting function) activity- is shown in red. The other two catalytic site residues Glu296 and Tyr383 are shown in blue.

5.4.4 Proteins using different residues in the same or overlapping site for different functions

Moonlighting proteins in this category are multi-domain proteins (listed in Table 5.4) which use overlapping functional sites for carrying out their primary and moonlighting function(s).

Table 5.4: Proteins using different residues in the same or overlapping site for different functions

Protein (Organism)	**Function 1**	**Function 2**	**Structure**	**Refs.**
Phosphoglucose isomerase *(O. cuniculus)*	Phosphoglucose isomerase (EC 5.3.1.9)	Autocrine motility factor, Neuroleukin, Differentiation & maturation mediator	1DQR, 1IAT	(Jeffery *et al.*, 2000; Read *et al.*, 2001)
Fructose-bisphosphate aldolase *(P. falciparum)*	Fructose-bisphosphate aldolase (EC 4.1.2.13)	Attaches actin to TRAP proteins	2PC4	(Bosch *et al.*, 2007)
Gpx4(*H. sapiens*)	Phospholipid hydroperoxide glutathione peroxidase (EC 1.11.1.12)	polymerised form has structural role in spermatozoa	2OBI	(Scheerer *et al.*, 2007)
S1O ribosomal/ protein *(E. coli)*	Component of ribosomal 30S subunit	Part of transcription anti-termination complex	1O9J	(Luo *et al.*, 2008)
Lens Crystallin/ Retinal DH *(E. edwardii)*	Lens Crystallin	Retinal DH (EC 1.2.1.3)	1O9J	(Bateman *et al.*, 2003)

5.4.4.1 Phosphoglucose isomerase (*Oryctolagus cuniculus, Mus musculus, Homo sapiens*)

Phosphoglucose isomerase (PGI, EC 5.3.1.9) is a glycolytic enzyme which catalyses the interconversion of glucose-6-phosphate and fructose-6-phosphate (Function 1). It is a known to moonlight as an autocrine motility factor (AMF, tumor-secreted cytokine that promotes cellular growth and motility), neuroleukin (a neurotrophic factor for neurons) and differentiation mediator in mammals (Function 2) (Jeffery *et al.*, 2000). The human PGI exists as a dimer comprising 3 domains: two (one large and one small) Rossmann fold domains and an α orthogonal bundle (Read *et al.*, 2001) (Figure 5.11). The known catalytic site residues are: Lys210, Glu216, Gly271, Arg272, Glu357, His388, Lys518 (1IAT). The PGI inhibitor erythrose 4-phosphate (E4P) is known to inhibit both the enzymatic and cell motility activities of PGI. Moreover, mutation of the catalytic residues resulted in significant reduction in the cell motility stimulating activity.

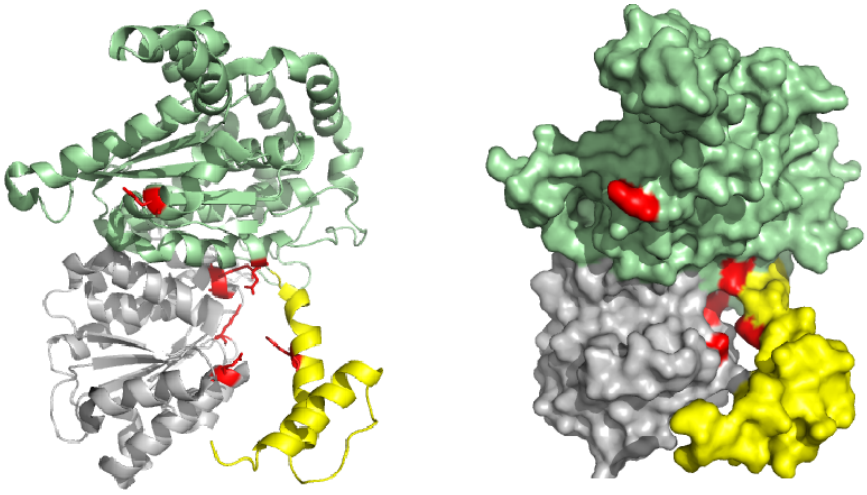


Figure 5.11: Primary and moonlighting functions of human Phosphoglucose isomerase (PGI)(PDB:1IAT). Catalytic site residues are shown as red sticks. Inhibition of enzymatic (primary function) and autocrine motility factor (moonlighting function) functions of PGI by the PGI inhibitor and mutational analysis of the catalytic residues have indicated overlapping regions of both functions in the human PGI.

5.4.4.2 Aldolase (*Plasmodium falciparum*)

The fructose-bisphosphate aldolase (EC 4.1.2.13; Function 1) from Apicomplexan parasites such as *P. falciparum* and *P. vivax* also provides a bridge between the actin filaments and TRAP (thrombospondin-related anonymous protein) which is critical for the host invasion machinery of the malaria parasite (Function 2) (Bosch *et al.*, 2007). The *P. falciparum* aldolase structure consists of 4 chains, each consisting of a TIM barrel domain (Figure 5.12). The aldolase active site residues and the residues involved in binding actin or TRAP overlap, are located in the centre of the TIM barrel. The aldolase active site consists of the residues Asp39, Lys112, Glu194 and Lys236 whilst the actin binding residues of aldolase are Arg48, Lys112, Arg153, Lys236 and the TRAP binding residues are Glu40, Lys47, Arg48, Lys151, Arg153, Arg309 and Gln312.

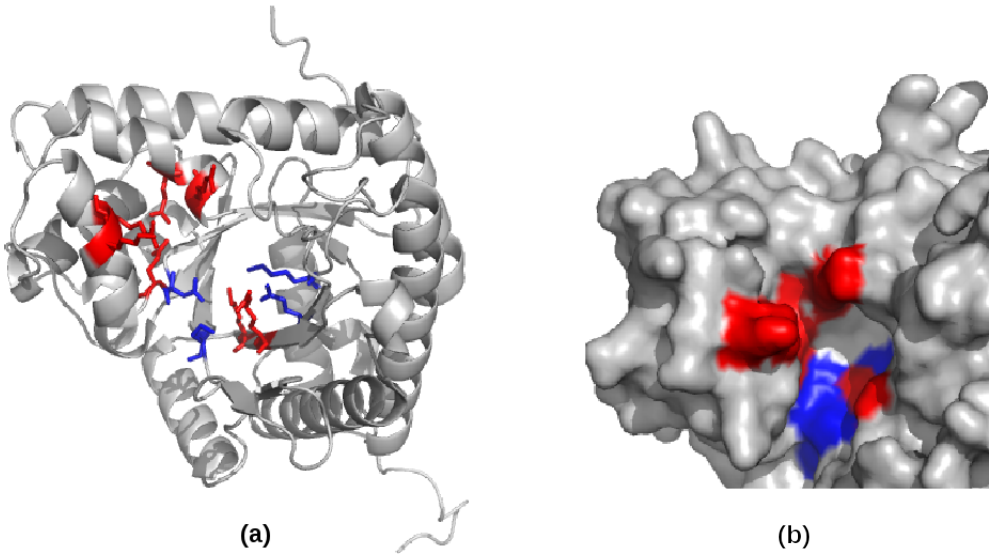


Figure 5.12: Primary and moonlighting functions of Aldolase. (a) Cartoon structure representation of *P. falciparum* aldolase (PDB: 2PC4) is shown. (b) This figure highlights the centre of the TIM barrel of the aldolase in surface representation. In both figures, the enzyme active site (primary function) is shown in blue and the actin-binding site (moonlighting function) is shown in red.

5.4.5 Proteins with different structural conformations for different functions

These are 'transformer proteins' (Knauer *et al.*, 2012) (listed in Table 5.5) which use different conformational states for carrying out their primary and moonlighting function(s).

Table 5.5: Proteins using different conformations for different functions

Protein (Organism)	Function 1	Function 2	Structure	Refs.
RfaH (*E. coli*)	Transcription factor	Translational regulator	2OUG, 2LCL	(Belogurov *et al.*, 2007) (Burmann *et al.*, 2012)
Lymphotactin (Ltn) *(H. sapiens)*	Chemokine (activates XCR1)	Binds cell-surface glycosaminoglycans	2JP1	(Tuinstra *et al.*, 2008)

5.4.5.1 RfaH (*E. coli*)

RfaH is a bacterial anti-termination protein which binds to the RNA polymerase (RNAP) and suppresses pausing, Rho-dependent inhibition and intrinsic termination at a subset of sites (Function 1) (Belogurov *et al.*, 2007). As a result, termination signals are bypassed, which allows complete synthesis of long RNA chains. RfaH is a two-domain protein. The two domains are observed to interact closely in the crystal structure. The RfaH N-terminal domain (NTD) has a central antiparallel β-sheet surrounded by α-helices and the C-terminal domain (CTD) in the crystal structure is an all--helical domain. However, the solution structure of the protein, solved by NMR, showed that the RfaH CTD folds into an α-helical structure when it interacts with the RfaH NTD and transforms into an all-β-sheet fold in the absence of NTD (Figure 5.13). These two different fold states allow the protein to perform alternative functions. When the CTD exists in the all-β-sheet state, it can stimulate translation by recruiting a ribosome to an mRNA lacking a ribosome-binding site (Function 2) (Burmann *et al.*, 2012).

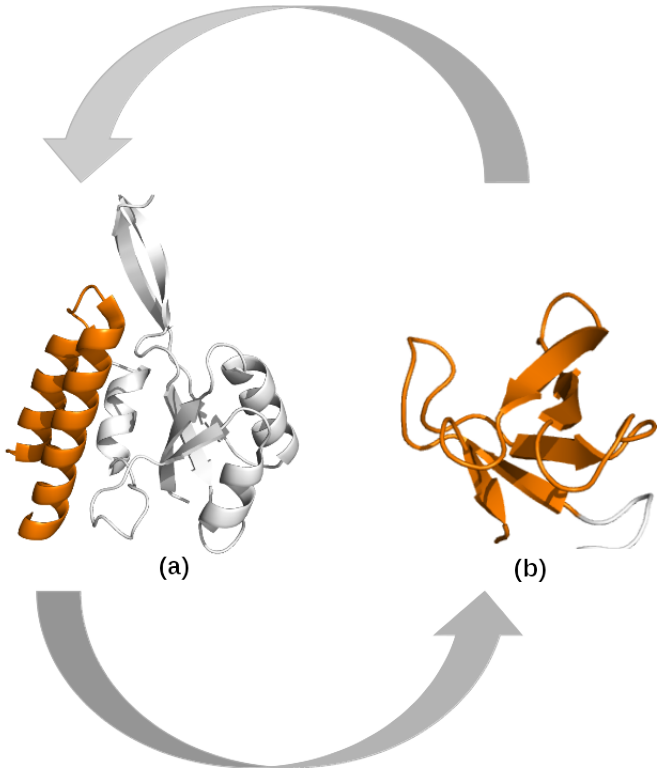


Figure 5.13: Primary and moonlighting functions of RfaH. The Rfah CTD is coloured in orange. In the closed form of RfaH (a), the CTD (α-helix form) and NTD tightly interacts, and it works as a transcription factor (PDB: 2OUG). The subsequent (or simultaneous) refolding of the CTD into a (b) β-barrel transforms RfaH into a translation factor (PDB: 2LCL).

5.5 Exploiting CATH FunFams to annotate moonlighting proteins

We used a dataset of 144 proteins from the database MultitaskProtDB (Hernández *et al.*, 2014) to analyse whether the functional annotations from CATH FunFams can be exploited to suggest moonlighting.

All analyses were performed on the UniProtKB/Swiss-Prot database and UniProtKB-GOA database (dated November 2013 and considering only non-IEA GO terms). The performance of FunFHMMer on the moonlighting protein dataset was benchmarked against PSI-BLAST, BLAST and Pfam families, since PSI-BLAST and Pfam were shown in previous studies (Gómez *et al.*, 2003; Khan *et al.*, 2012; Hernández *et al.*, 2015) to perform well in predicting the moonlighting functions

of proteins.

PSI-BLAST was performed with the default setting of three iterations. Then all hits with an E-value score < 0.01 that have annotations, were used for transferring annotations to the query sequence. The GO term predictions were labelled according to the annotation frequency of a particular GO term amongst the PSI-BLAST hits and propagated up the GO directed acyclic graph (DAG). For the Pfam and FunFHMMer predictions, the moonlighting predictions were removed from the seed sequences of the respective Pfam families or CATH FunFams and their corresponding HMMs were then generated. The moonlighting proteins were then scanned against the HMMs and the GO terms of their FunFam top hits (E-value < 0.01) were transferred to the query in a probabilistic manner calculated as the annotation frequency in a matched family and propagated up the GO DAG.

Performance of function predictions made by FunFHMMer compared with PSI-

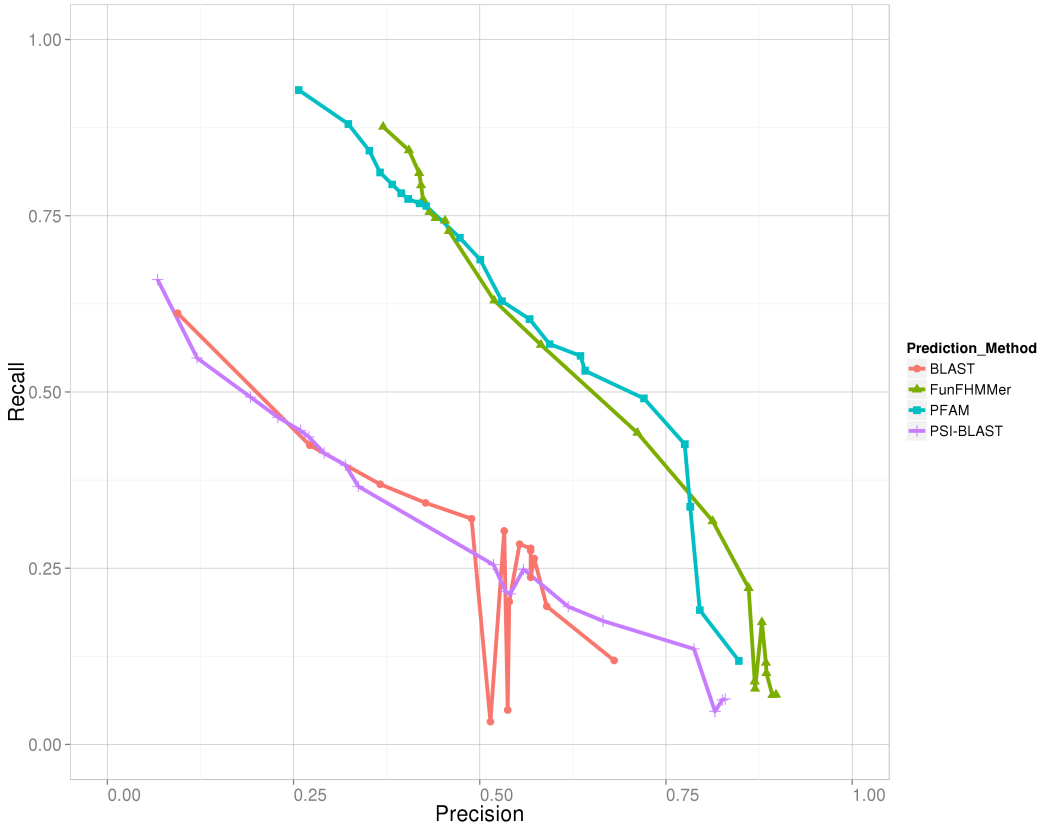


Figure 5.14: Comparison of the performance of FunFHMMer with PSI-BLAST, BLAST and Pfam in prediction of moonlighting proteins.

BLAST (number of iterations = 3), BLAST and Pfam is illustrated in Figure 5.14 for Molecular Function Ontology (MFO) using a precision-recall (*pr-rc*) curve (see Section 3.1.3.1 in Chapter 3). The figure clearly indicates that both FunFHMMer and Pfam perform competitively and better than both BLAST and PSI-BLAST in predicting GO terms for the 144 moonlighting proteins in the dataset.

As mentioned before, methods aiming to detect diverse sequences (i.e. PSI-BLAST, PFP,ESG, or scans of Pfam families) can help in capturing the functional diversity of moonlighting proteins and aid in predicting secondary or alternative functions of these proteins, as these alternative functions are sometimes present in remote homologues (Khan *et al.*, 2012; Hernández *et al.*, 2015). However, the FunFHMMer protocol is designed to predict functions based on functionally coherent FunFams. This would be expected to distinguish between relatives with alternative functions when these are associated with different sequence motifs.

For example, the Chaperonin 60 apical domain (CATH 3.50.7.10) sequences for *Homo sapiens* and *Enterobacter aerogenes* which have two different moonlighting functions (Henderson *et al.*, 2013) are split into two different FunFams (3979 and 3904 respectively) in CATH v4.0 FunFams for the apical domain superfamily. Moreover, an analysis of the conserved residues of the FunFams showed that FunFHMMer had identified the moonlighting motif which was reported in the literature (see Figure 5.15).

Figure 5.15: Partial sequence logos of CATH FunFams 3904 and 3979 is shown that is generated using WebLogo 3.0 (Crooks *et al.*, 2004). The known moonlighting motif (in green) in human HSP60 sequence is highly conserved in its best match family in CATH-Gene3D (FunFam 3904) in the Chaperonin 60 apical domain superfamily but it is absent in a closely related family (FunFam 3979) containing bacterial sequences which have a different moonlighting activity.

5.6 Conclusion and Discussion

From the detailed analysis of the moonlighting proteins examined above, one can see considerable structural diversity in the types of domains that have evolved a moonlighting function together with significant diversity in the different types of moonlighting functions that have evolved in these proteins. Some of the moonlighting proteins utilize different sites for their primary and moonlighting functions, however, there are others which use overlapping regions with their primary functional site or even the same site for both functions. Our investigation of moonlighting proteins revealed two general trends of functional site utilization in moonlighting proteins:

(i) **Type 1:** The primary functional site resides in the largest pocket of the protein while the moonlighting functional site is present on a distinct exposed surface of the protein. This is true for proteins for which binding to other proteins facilitates their moonlighting function. Examples: Enolase, Perodoxin, MAPK1, PutA, I-Anil

(ii) **Type 2:** The primary and moonlighting functional sites are present on two pockets or clefts in the protein structure. These sites can be utilized for two different enzymatic reactions or an enzymatic reaction together with a binding function. Examples: Cytochrome c, Albaflavenone Monooxygenase.

Figure 5.16 shows the structural and functional diversity of the CATH domain superfamilies that are represented in the moonlighting proteins discussed in this chapter. We observe that these proteins belong to superfamilies ranging from very low to high structural and functional diversity. Knowledge of the structural and functional diversity sheds some light on the possible routes by which they may have acquired their moonlighting function. For example, for domain superfamilies with high structural diversity, it is more likely that the new functions can emerge through structural embellishments. By contrast, domain superfamilies having low structural diversity are more likely to evolve a new function by domain recruitment

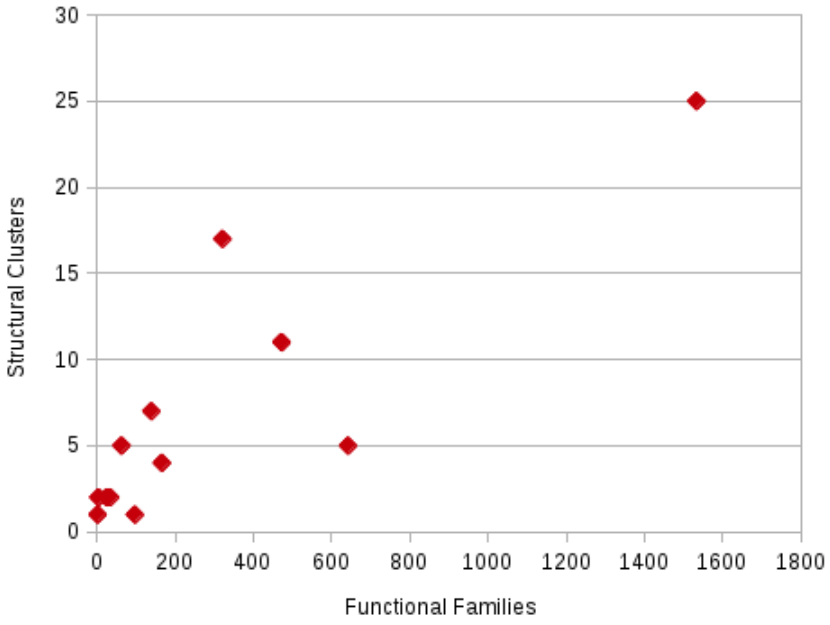


Figure 5.16: Structural diversity vs functional diversity of CATH superfamilies represented in the moonlighting protein dataset studied in this chapter. Structural diversity is represented by the number of structural clusters (domains clustered at 5ÅRMSD) in the superfamilies and the functional diversity is represented by the number of functional families identified in the superfamily.

or use of different amino acids.

Function annotations were predicted using the FunFHMMer function prediction protocol and compared to predictions made by PSI-BLAST, BLAST and Pfam. FunFHMMer outperformed predictions from both BLAST and PSI-BLAST and performed competitively with annotations predicted from Pfam families.

Previously, studies by Gómez *et al.* (2003); Khan *et al.* (2012) have established that the inference of functional annotations from remote homologs using PSI-BLAST, ESG and PFP is useful in identifying the moonlighting or alternative functions of a protein. In practice these methods can be viewed as searching for a 'needle in a haystack' as it is difficult for a researcher to identify the correct hit from the large output from these tools. However, analysis of the results of the moonlighting dataset benchmark described in our work shows that functions predicted by protein families such as Pfam and the CATH FunFams outperforms PSI-BLAST. This suggests that an alternative sequence-based approach to iden-

tify the moonlighting or alternative functions of a protein is to use a finer classification of close homologs such as Pfam or CATH FunFams to identify moonlighting motifs that can aid in identifying the moonlighting function of proteins. Moreover, the results of this approach would be easier for researchers to interpret.

Analysis of the known moonlighting proteins in terms of functionality, physical locations in the cell, mechanisms to moonlight and the type of genomes they are found to exist in, sows that moonlighting proteins are diverse in nature. Jeffery (2015) suggests that moonlighting may be a more common phenomenon in proteins than currently thought. This is supported by the view that moonlighting provides the cell with an economical strategy to re-utilise or re-purpose existing proteins for alternative purposes by avoiding synthesis of new proteins (Royle, 2013). However, only $\sim$ 300 moonlighting proteins are currently known, $<$ 40% of these have known structures and only a handful of the proteins having structure have experimentally characterised functional site information. As the number of moonlighting proteins with experimental characterisations and structure information increases, these analyses can be pursued further.

Chapter 6

Conclusions and Future directions

6.1 Summary of work

Analysis of protein function using a domain grammar of function requires functional classification of a protein domain resource into coherent functional groups. Towards this end, protein domain sequences in CATH-Gene3D superfamilies were originally classified into functional families or FunFams by partitioning a hierarchical tree of sequence relatives for each superfamily, produced by an in-house agglomerative hierarchical clustering method, at a generic threshold (Lee *et al.*, 2010). This classification was later improved by a family identification method which determined the optimal partitioning of the superfamily clustering tree by exploiting functional annotation data from the Gene Ontology (GO) to sub-classify the domain superfamilies into FunFams (Rentzsch and Orengo, 2013).

In this work, a new and improved method, FunFHMMer, for functionally classifying CATH domain superfamilies was designed and developed. For each superfamily in CATH, FunFHMMer identifies functional families or FunFams by determining an optimal partitioning of the superfamily hierarchical clustering tree on the basis of sequence information entirely and hence, is unaffected by limitations of function annotation resources.

The first work chapter of this thesis describes the development of the FunFHMMer algorithm. FunFHMMer analyses sequence alignments of all parent node clusters in the superfamily tree and identifies specificity-determining positions and conserved positions using GroupSim (Capra *et al.*, 2009). Various sequence-based parameters were assessed to determine those that are critical for inferring functional coherence of alignments by an in-depth analysis of the large, well-studied and diverse Thiamine pyrophosphate (TPP)-dependent enzyme superfamily. These parameters were incorporated in FunFHMMer to calculate a novel

index which is used to assess functional coherence of a parent node. The superfamily tree is then partitioned at the parent nodes that are inferred as not being functionally coherent. Functional classification of 2735 protein domain superfamilies in CATH (version 4.0) by FunFHMMer resulted in the generation of 110,439 FunFams that are significantly more functionally pure than in the previous classifications. The functional annotations provided by FunFams were found to be more precise compared with those generated by other domain-based resources.

In the second work chapter, the predictive power of the FunFams was examined and it was found to provide better performance in function prediction for uncharacterised sequences compared to other domain-based resources using a UniProtKB/Swiss-Prot rollback assessment. Furthermore, the function prediction pipeline based on the FunFams was ranked among the top 10 function prediction methods in predicting GO terms in the Critical Assessment of protein Function Annotation (CAFA) 2 international function prediction experiment. A web server was set up to make the function prediction pipeline publicly-available.

In the third work chapter, an in-depth analysis of the FunFams generated by FunFHMMer is performed which aids in the identification of steps that may improve the quality of the FunFams. The utility of the FunFams to explore superfamily diversity was examined and it was found that the FunFams clearly capture information on structural, functional (measured by EC number annotations) and multi-domain architecture diversity within superfamilies. The ability of the sequence conservation information in the FunFams in capturing protein functional site information was manually assessed using the serine beta-lactamases which are implicated in antibiotic resistance. Detailed analysis of the functional determinants identified in serine beta-lactamases by the FunFams i.e. residues differing between three serine beta-lactamase Class FunFams showed that these residue positions were likely to contribute to the different implementations of the catalytic mechanism in the three Classes. Following this, a functional site pre-

diction method, FunSite, was developed that predicts structural clusters of highly conserved residues of a FunFam as functional sites for sequences assigned to the FunFam. An analysis of 246 protein domains showed that the performance of FunSite is competitive to the widely-used Evolutionary Trace (ET) method.

In the last work chapter, the structure-function relationships of moonlighting proteins was examined and a classification of these proteins was proposed thereafter. The ability of FunFams was also assessed for providing functional annotations for moonlighting proteins and its performance was found to be competitive with that of Pfam, both of which outperformed PSI-BLAST, which had been previously reported to perform better than other family-based or domain-based resources.

In summary, a new method for functional classification of protein domain superfamilies in CATH into FunFams was developed. The FunFams were found to be functionally pure and their use in protein function annotation and prediction of functional sites using a domain-centric approach was validated by known functional information. Furthermore, the FunFams provided a powerful tool in studying evolution of function in protein domains.

6.2 Future directions

Whilst the FunFams in the CATH-Gene3D resource have been validated to be reasonably effective in transferring experimental annotations between relatives, there is still considerable room for improvement. The FunFams are still a coarse level of clustering that may be improved by incorporating other parameters such as structural data and multi-domain architecture information. For example, for the analysis of serine beta-lactamases the FunFams were able to differentiate between the three Classes (A, C and D) of serine beta-lactamases (see Section 4.5.1 in Chapter 4), however, in order to differentiate between different types of Class A beta-lactamases that confer different phenotypes i.e. antibiotic resis-

tance, a more classification-based approach exploiting structural data called the Active Site Structural Profile (ASSP) protocol was developed by Dr. David Lee (Lee et al., 2016). ASSP constructs an active site profile of a representative structure of a FunFam by considering a 8Å radius around a known catalytic residue. The active site profile residues are then mapped to associated sequences in the FunFam and the FunFam is then sub-classified into subfamilies by clustering at 60% sequence identity. ASSP then determines the conserved residues and functional determinants in different subfamilies by a parsimony analysis of the residues in the active site profile region in all sequences of the subfamilies of the FunFam.

In future, it should be possible to improve the quality of the FunFams by making improvements to the FunFHMMer protocol in different aspects. A few potential changes to the FunFHMMer protocol are discussed in the following sections.

6.2.1 Use of structural data

The FunFHMMer functional classification protocol currently focuses on sequence data. Use of structural data along with the sequence conservation information in the FunFams could help in the identification of likely functional sites that could assist in assessing the functional coherence of sequence clusters by focussing on likely functional sites. This could help to improve the functional quality of FunFams by preventing the merging of sequence relatives that do not share the same functional sites. A probable automated strategy to improve the functional purity of the CATH FunFams is illustrated in Figure 6.1 that uses conserved site information in the FunFams and structural data (where available) to focus on likely functional sites.

This would first involve residue conservation analysis of FunFams using Scorecons. The conserved residue data along with structural data available for the sequence relatives of a FunFam would be used for prediction of functional sites in

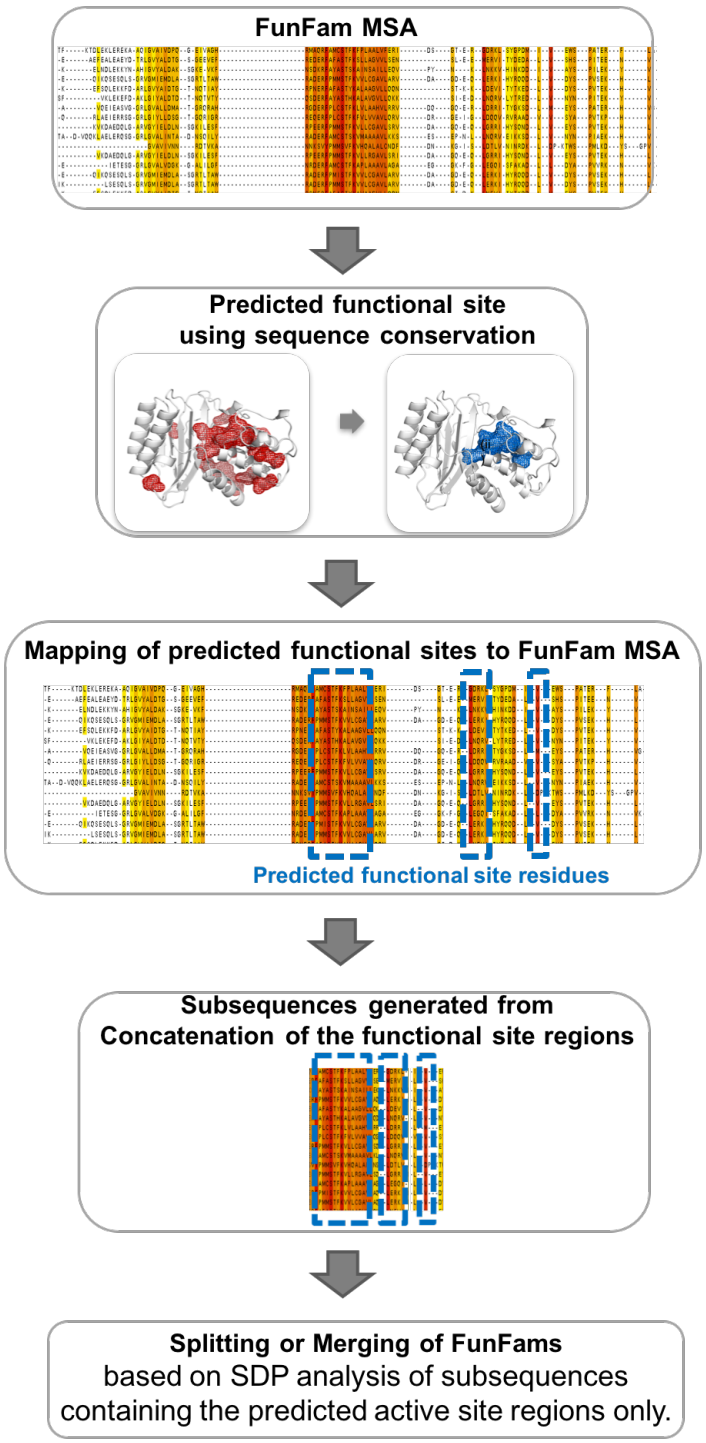


Figure 6.1: A strategy to improve the functional purity of CATH FunFams.

the FunFam. These predicted functional site residues in FunFams would then be used to infer whether sequence relatives in the FunFam share the same functional sites or not. This information would not only help in improving the functional purity of a large number of FunFams by splitting them based on the similarity of

functional sites in sequence relatives but will also help in merging FunFams within a superfamily that share the same functional sites. This could help to reduce the number of FunFams having low information content and improve the quality the FunFams in CATH. Alternatively, this strategy could also be integrated into the FunFHMMer protocol during generation of FunFams to improve the assessment of functional coherence of sequence clusters by focussing on predicted functional sites.

6.2.2 Changes to GeMMA

The CATH superfamily clustering trees generated by the GeMMA clustering algorithm (Lee *et al.*, 2010) guides and limits the pairwise sequence cluster comparisons of the FunFHMMer protocol (Das *et al.*, 2015b). The high-throughput version of GeMMA (see Section 2.1.4.1) implements two heuristics - greedy merging and comparison sampling - that have been reported to maintain the same performance levels as the original GeMMA clustering protocol (Lee *et al.*, 2010). However, the heuristics may have potential impacts (see Figure 6.2) on the clustering process which may affect the quality of FunFams generated by the FunFHMMer pipeline (Rentzsch, 2012).

It should be possible to improve the performance of FunFHMMer by either (1) only applying changes to the GeMMA heuristics to certain stages of clustering depending on the size of the superfamily sequence dataset or (2) replacing the GeMMA clustering method by a better-performing clustering approach. Furthermore, the GeMMA clustering method used in the FunFHMMer protocol may be changed or replaced by a better-performing clustering method to give a more accurate clustering tree of sequence relatives for CATH superfamilies.

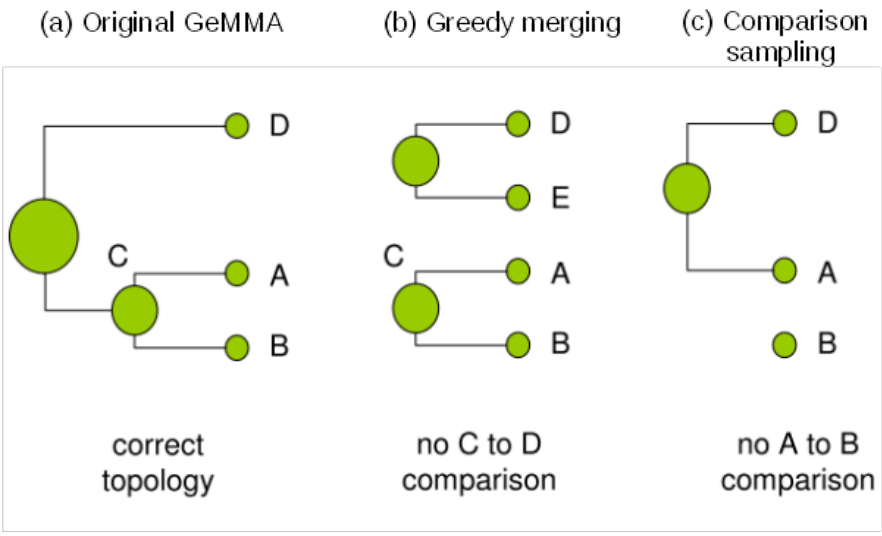


Figure 6.2: Potential impact of the GeMMA heuristics on the clustering tree. The green circles represent the node sequence clusters. **(a)** The original GeMMA tree is generated by all-against-all cluster comparisons where only the most closely related pair of clusters are merged at each iteration. **(b)** Use of greedy merging heuristics results in merging of all clusters that meet a particular similarity threshold in each iteration. This can result in a clustering tree where clusters C and D are never compared as D gets merged with E at the same iteration in which C is created. **(c)** Use of comparison sampling heuristics results in a randomly drawn subset of comparisons to be carried out in each iteration. This can result in cases where the pair A and B are not drawn in the first iteration, and A is then merged with D. In the next iteration, A and B are not compared and A is already merged to another cluster. Adapted from (Rentzsch, 2012).

6.3 Final remarks

Our knowledge of the protein repertoire is expanding rapidly as the international genomics initiatives and metagenomics initiatives continue. However, less than 1% of these proteins have experimentally characterised function annotations (UniProt-Consortium, 2015). Clinical data is also accumulating, which links genetic variations in proteins with disease. To make sense of all this data, computational approaches are required to predict the functions of the proteins identified and determine the location of residue sites that are essential for these functions.

When an uncharacterised protein does not show sufficient similarity to any characterised whole protein, its function can perhaps be better understood by analysing its domain components and finding functionally characterised homologs

for each domain. Exploiting this approach, a domain grammar (Dessailly *et al.*, 2009) can be used to describe protein function. The success of domain-based strategies (Fang and Gough, 2013; Rentzsch and Orengo, 2013; Das *et al.*, 2015b) for protein function prediction in CAFA (Radivojac *et al.*, 2013; Jiang *et al.*, 2016) also suggests that there is considerable signal in the domain, reflecting the proteins molecular function and the context in which it operates.

The search for homologs using domain functional families can not only help in improving the reliability of function predictions for uncharacterised sequences, but it can also aid in the identification of highly conserved features of a functional family, which are expected to be functionally important. This ability to assign uncharacterised sequences to functional families and obtain information on conserved functional sites will be important for understanding the consequence of residue mutations in genetic variants of these proteins. Furthermore, it potentially allows more sensitive detection of new family members by homology recognition, better discrimination between non-members and identification of novel sequences that do not match any existing families.

This work describes a new approach for functional sub-classification of the CATH-Gene3D domain superfamilies into functional families or (FunFams) based on the difference in specificity-determining positions between functional groups. This results in a domain classification that is able to provide accurate functional annotations than broader groupings of relatives such as the previous functional classification in CATH (i.e. families generated by DFX algorithm), Pfam and CDD families. The predictive power of the FunFams was validated by their performance in the recent CAFA 2 experiment (Jiang *et al.*, 2016). Furthermore, their utility in exploring functional diversity at the domain-level (Das *et al.*, 2015a), annotating metagenome data (Dawson, 2015), providing good templates for homology modelling (Lam *et al.*, 2016), and identifying new drug targets (Garcia *et al.*, 2016) was also demonstrated recently. It is hoped that with the accumulation of

more sequences, structures and function annotations, the domain-based functional classification developed by this project will provide more accurate function annotations and increased coverage in annotating uncharacterised proteins and that the resulting functional family profiles will provide useful information towards making sense of the vast uncharacterised sequence and structural data available.

References

Abhiman, S. and Sonnhammer, E. L. (2005). FunShift: a database of function shift analysis on protein subfamilies, *Nucleic Acids Research*, **33**.suppl 1, D197–D200. 39

Addou, S., Rentzsch, R., Lee, D. and Orengo, C. A. (2009). Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer, *Journal of Molecular Biology*, **387**.2, 416–430. 103, 104

Aguilera, L., Giménez, R., Badia, J., Aguilar, J. and Baldoma, L. (2010). NAD+-dependent post-translational modification of *Escherichia coli* glyceraldehyde-3-phosphate dehydrogenase, *International Microbiology*, **12**.3, 187–192. 213, 214

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. and Dessimoz, C. (2012). Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs, *PLoS Computational Biology*, **8**.5, e1002514. 32

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool, *Journal of Molecular Biology*, **215**.3, 403–410. 41, 62, 104, 129

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25**.17, 3389–3402. 104, 202

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. and Murzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining, *Nucleic Acids Research*, **42**.D1, D310–D314. 50

Anstrom, D. M. and Remington, S. J. (2006). The product complex of *M. tuberculosis* malate synthase revisited, *Protein science*, **15**.8, 2002–2007. 209, 211

Appleton, B. A., Wu, P., Maloney, J., Yin, J., Liang, W.-C., Stawicki, S., Mortara, K., Bowman, K. K., Elliott, J. M., Desmarais, W. et al. (2007). Structural studies of neuropilin/antibody complexes provide insights into semaphorin and VEGF binding, *The EMBO journal*, **26**.23, 4902–4912. 210

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000). Gene Ontology: tool for the unification of biology, *Nature Genetics*, **25**.1, 25–29. 25, 26, 70

Ashkenazy, H., Erez, E., Martz, E., Pupko, T. and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, *Nucleic Acids Research*, **38**.suppl 2, W529–W533. 151, 154

Atkinson, H. J., Morris, J. H., Ferrin, T. E. and Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies, *PloS One*, **4**.2, e4345–e4345. 170

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., Roma-Mateo, C., Theodosiou, A. and Mitchell, A. L. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012, *Database*, **2012**, bas019. 46

Bader, G. D., Betel, D. and Hogue, C. W. (2003). BIND: the biomolecular interaction network database, *Nucleic Acids Research*, **31**.1, 248–250. 202

Baier, F., Chen, J., Solomonson, M., Strynadka, N. C. and Tokuriki, N. (2015). Distinct metal isoforms underlie promiscuous activity profiles of metalloenzymes, *ACS Chemical Biology*, **10**.7, 1684–1693. 59

Baier, F. and Tokuriki, N. (2014). Connectivity between catalytic landscapes of the metallo-β-lactamase superfamily, *Journal of Molecular Biology*, **426**.13, 2442–2456. 59

Bairoch, A. (1994). The ENZYME data bank, *Nucleic Acids Research*, **22**.17, 3626–3627. 25, 26

Bairoch, A. (2000). The ENZYME database in 2000, *Nucleic Acids Research*, **28**.1, 304–305. 70, 100

Bartlett, G. J., Porter, C. T., Borkakoti, N. and Thornton, J. M. (2002a). Analysis of Catalytic Residues in Enzyme Active Sites, *Journal of Molecular Biology*, **324**.1, 105 – 121. 38, 82, 185

Bartlett, G. J., Porter, C. T., Borkakoti, N. and Thornton, J. M. (2002b). Analysis of catalytic residues in enzyme active sites, *Journal of Molecular Biology*, **324**.1, 105–121. 154, 186

Bashton, M. and Chothia, C. (2007). The generation of new protein functions by the combination of domains, *Structure*, **15**.1, 85–99. 25, 109

Bateman, O., Purkiss, A., Van Montfort, R., Slingsby, C., Graham, C. and Wistow, G. (2003). Crystal structure of η-crystallin: adaptation of a class 1 aldehyde dehydrogenase for a new role in the eye lens, *Biochemistry*, **42**.15, 4349–4356. 217

Baumgartner, W. A., Cohen, K. B., Fox, L. M., Acquaah-Mensah, G. and Hunter, L. (2007). Manual curation is not sufficient for annotation of genomic databases, *Bioinformatics*, **23**.13, i41–i48. 102

Beadle, G. W. and Tatum, E. L. (1941). Genetic control of biochemical reactions in *Neurospora*, *Proceedings of the National Academy of Sciences*, **27**.11, 499–506. 199

Becker, E., Robisson, B., Chapple, C. E., Guénoche, A. and Brun, C. (2012). Multifunctional proteins revealed by overlapping clustering in protein interaction network, *Bioinformatics*, **28**.1, 84–90. 202

Belogurov, G. A., Vassylyeva, M. N., Svetlov, V., Klyuyev, S., Grishin, N. V., Vassylyev, D. G. and Artsimovitch, I. (2007). Structural basis for converting a general transcription factor into an operon-specific virulence regulator, *Molecular Cell*, **26**.1, 117–129. 220

Bhoumik, A., Takahashi, S., Breitweiser, W., Shiloh, Y., Jones, N. and Ronai, Z. (2005). ATM-dependent phosphorylation of ATF2 is required for the DNA damage response, *Molecular Cell*, **18**.5, 577–587. 210

Binkowski, T. A., Freeman, P. and Liang, J. (2004). pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins, *Nucleic Acids Research*, **32**.suppl 2, W555–W558. 108

Bosch, J., Buscaglia, C. A., Krumm, B., Ingason, B. P., Lucas, R., Roach, C., Cardozo, T., Nussenzweig, V. and Hol, W. G. (2007). Aldolase provides an unusual binding site for thrombospondin-related anonymous protein in the invasion machinery of the malaria parasite, *Proceedings of the National Academy of Sciences*, **104**.17, 7015–7020. 217, 219

Brady Jr, G. P. and Stouten, P. F. (2000). Fast prediction and visualization of protein binding pockets with pass, *Journal of computer-aided molecular design*, **14**.4, 383–401. 156

Brenner, S. E. (1999). Errors in genome annotation, *Trends in Genetics*, **15**.4, 132–133. 119

Brown, D. P., Krishnamurthy, N. and Sjölander, K. (2007). Automated protein subfamily identification and classification, *PLoS Computational Biology*, **3**.8, e160. 67, 68, 93

Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K. and Haussler, D., (1993). Using dirichlet mixture priors to derive hidden markov models for protein families. In *Ismb*, volume 1, pages 47–55. 66

Brown, S. D. and Babbitt, P. C. (2014). New insights about enzyme evolution from large scale studies of sequence and structure relationships, *Journal of Biological Chemistry*, **289**.44, 30221–30228. 52, 95, 153, 170

Brown, S. D., Gerlt, J. A., Seffernick, J. L. and Babbitt, P. C. (2006). A gold standard set of mechanistically diverse enzyme superfamilies, *Genome biology*, **7**.1, R8. 67, 68

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D, *Nucleic Acids Research*, **33**.suppl 1, D212–D215. 105

Brylinski, M. and Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation, *Proceedings of the National Academy of sciences*, **105**.1, 129–134. 157

Burmann, B. M., Knauer, S. H., Sevostyanova, A., Schweimer, K., Mooney, R. A., Landick, R., Artsimovitch, I. and Rösch, P. (2012). An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor, *Cell*, **150**.2, 291–303. 220

Cammer, S. A., Hoffman, B. T., Speir, J. A., Canady, M. A., Nelson, M. R., Knutson, S., Gallina, M., Baxter, S. M. and Fetrow, J. S. (2003). Structure-based active site profiles for genome analysis and functional family subclassification, *Journal of Molecular Biology*, **334**.3, 387–401. 151

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) database: sharing knowledge in UniProt with Gene Ontology, *Nucleic Acids Research*, **32**.suppl 1, D262–D266. 105

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics, *Nucleic Acids Research*, **37**.suppl 1, D233–D238. 68

Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M. and Funkhouser, T. A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, *PLoS Computational Biology*, **5**.12, e1000585. 151, 154, 156, 227

Capra, J. A. and Singh, M. (2008). Characterization and prediction of residues determining protein functional specificity, *Bioinformatics*, **24**.13, 1473–1480. 40, 77, 82, 92, 93, 177, 178, 198

Casari, G., Sander, C. and Valencia, A. (1995). A method to predict functional residues in proteins, *Nature structural biology*, **2**.2, 171. 39

Chagoyen, M., García-Martín, J. A. and Pazos, F. (2015). Practical analysis of specificity-determining residues in protein families, *Briefings in Bioinformatics*, **33**.14, 4455–4465. 38

Chakrabarti, S., Bryant, S. H. and Panchenko, A. R. (2007). Functional specificity lies within the properties and evolutionary changes of amino acids, *Journal of Molecular Biology*, **373**.3, 801–810. 78

Chakraborty, A. and Chakrabarti, S. (2015). A survey on prediction of specificity-determining sites in proteins, *Briefings in Bioinformatics*, **16**.1, 71–88. 40, 77, 78

Chen, C.-W., Lin, J. and Chu, Y.-W. (2013). iStable: off-the-shelf predictor integration for predicting protein stability changes, *BMC Bioinformatics*, **14**.2, 1. 185

Chen, H. and Zhou, H.-X. (2005). Prediction of solvent accessibility and sites of deleterious mutations from protein sequence, *Nucleic Acids Research*, **33**.10, 3193–3199. 185

Cheng, H., Liao, Y., Schaeffer, R. D. and Grishin, N. V. (2015). Manual classification strategies in the ECOD database, *Proteins: Structure, Function, and Bioinformatics*, **83**.7, 1238–1251. 51

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., Kim, B.-H. and Grishin, N. V. (2014). ECOD: an evolutionary classification of protein domains, *PLoS Computational Biology*, **10**.12, e1003926. 51

Chitale, M., Hawkins, T., Park, C. and Kihara, D. (2009). ESG: extended similarity group method for automated protein function prediction, *Bioinformatics*, **25**.14, 1739–1745. 202

Chitale, M., Khan, I. K. and Kihara, D. (2013). In-depth performance evaluation of PFP and ESG sequence-based function prediction methods in CAFA 2011 experiment, *BMC Bioinformatics*, **14**.suppl 3, S2. 117

Chothia, C., Gough, J., Vogel, C. and Teichmann, S. A. (2003). Evolution of the protein repertoire, *Science*, **300**.5626, 1701–1703. 25

Chung, S. Y. and Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology, *Structure*, **4**.10, 1123–1127. 109

Clark, W. T. and Radivojac, P. (2011). Analysis of protein function and its prediction from amino acid sequence, *Proteins: Structure, Function, and Bioinformatics*, **79**.7, 2086–2096. 109

Clark, W. T. and Radivojac, P. (2013). Information-theoretic evaluation of predicted ontological annotations, *Bioinformatics*, **29**.13, i53–i61. 116

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering, *Nucleic acids research*, **16**.22, 10881–10890. 71

Costa, E. P., Vens, C. and Blockeel, H. (2013). Top-down clustering for protein subfamily identification, *Evolutionary bioinformatics online*, **9**, 185. 77

Costelloe, S. J., Ward, J. M. and Dalby, P. A. (2008). Evolutionary analysis of the TPP-dependent enzyme family, *Journal of Molecular Evolution*, **66**.1, 36–49. 57, 76, 146

Cozzetto, D., Buchan, D. W., Bryson, K. and Jones, D. T. (2013). Protein function prediction by massive integration of evolutionary analyses and multiple data sources, *BMC Bioinformatics*, **14**.suppl 3, S1. 109, 110, 117

Crennell, S., Takimoto, T., Portner, A. and Taylor, G. (2000). Crystal structure of the multifunctional paramyxovirus hemagglutinin-neuraminidase, *Nature Structural & Molecular Biology*, **7**.11, 1068–1074. 213

Crooks, G. E., Hon, G., Chandonia, J.-M. and Brenner, S. E. (2004). Weblogo: a sequence logo generator, *Genome research*, **14**.6, 1188–1190. 93, 223

Cuff, A., Redfern, O. C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T., Todd, A. et al. (2009). The CATH hierarchy revisited–structural divergence in domain superfamilies and the continuity of fold space, *Structure*, **17**.8, 1051–1062. 52, 53, 70, 167

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S. et al. (2015). Ensembl 2015, *Nucleic Acids Research*, **43**.D1, D662–D669. 50

Das, S., Dawson, N. L. and Orengo, C. A. (2015a). Diversity in protein domain superfamilies, *Current Opinion in Genetics & Development*, **35**, 40–49. 52, 53, 54, 58, 70, 234

Das, S., Lee, D., Sillitoe, I., Dawson, N. L., Lees, J. G. and Orengo, C. A. (2015b). Functional classification of CATH superfamilies: a domain-based approach for protein function annotation, *Bioinformatics*, **31**.21, 3460–3467. 105, 106, 151, 182, 196, 232, 234

Das, S. and Orengo, C. A. (2016). Protein function annotation using protein domain family resources, *Methods*, **93**, 24–34. 13, 73, 78, 81, 84, 87, 91, 95, 97, 99, 122, 127

Das, S., Sillitoe, I., Lee, D., Lees, J. G., Dawson, N. L., Ward, J. and Orengo, C. A. (2015). CATH FunFHMMer web server: protein functional annotations using functional family assignments, *Nucleic Acids Research*, page gkv488. 105

Dawson, N. L., (2015). *Characterising functional diversity in protein domain superfamilies and metagenomes*. PhD thesis, University College London. 100, 234

Dayhoff, M. O. and Schwartz, R. M., (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*. 33

de Beer, T. A., Berka, K., Thornton, J. M. and Laskowski, R. A. (2014). PDBsum additions, *Nucleic Acids Research*, **42**.D1, D292–D296. 44, 204

de Lima Morais, D. A., Fang, H., Rackham, O. J., Wilson, D., Pethica, R., Chothia, C. and Gough, J. (2010). SUPERFAMILY 1.75 including a domain-centric gene ontology method, *Nucleic Acids Research*, **39**.suppl 1, D427–D434. 50

del Sol Mesa, A., Pazos, F. and Valencia, A. (2003). Automatic methods for predicting functionally important residues, *Journal of Molecular Biology*, **326**.4, 1289–1302. 151

Dessailly, B. H., Dawson, N. L., Mizuguchi, K. and Orengo, C. A. (2013). Functional site plasticity in domain superfamilies, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, **1834**.5, 879–889. 56, 82, 98, 153, 154

Dessailly, B. H., Redfern, O. C., Cuff, A. and Orengo, C. A. (2009). Exploiting structural classifications for function prediction: towards a domain grammar for protein function, *Current Opinion in Structural Biology*, **19**.3, 349–356. 234

Dessailly, B. H., Redfern, O. C., Cuff, A. L. and Orengo, C. A. (2010). Detailed analysis of function divergence in a large and diverse domain superfamily: toward a refined protocol of function classification, *Structure*, **18**.11, 1522–1535. 54

Dessimoz, C., Škunca, N. and Thomas, P. D. (2013). CAFA and the open world of protein function predictions, *Trends in genetics: TIG*, **29**.11, 609–610. 119

Devos, D. and Valencia, A. (2000). Practical limits of function prediction, *Proteins: Structure, Function, and Bioinformatics*, **41**.1, 98–107. 102, 103, 109

Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation, *TRENDS in Genetics*, **17**.8, 429–431. 109

Dimmer, E. C., Huntley, R. P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M. J., Bely, B., Browne, P., Chan, W. M., Eberhardt, R. et al. (2012). The uniprot-go annotation database in 2011, *Nucleic acids research*, **40**.D1, D565–D570. 74, 79

Do, C. B., Mahabhashyam, M. S., Brudno, M. and Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Research*, **15**.2, 330–340. 36

Duggleby, R. G. (2006). Domain relationships in thiamine diphosphate-dependent enzymes, *Accounts of chemical research*, **39**.8, 550–557. 76

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y. and Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues, *Nucleic Acids Research*, **34**.suppl 2, W116–W118. 108

Eddy, S. R. (1998). Profile hidden Markov models, *Bioinformatics*, **14**.9, 755–763. 42, 43

Eddy, S. R., (2009). A new generation of homology search tools based on probabilistic inference. In *Genome Inform*, volume 23, pages 205–211. World Scientific. 43, 76, 90, 121, 125, 146, 148, 177, 183

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput, *Nucleic acids research*, **32**.5, 1792–1797. 36

Ehinger, S., Schubert, W.-D., Bergmann, S., Hammerschmidt, S. and Heinz, D. W. (2004). Plasmin (ogen)-binding α-enolase from *Streptococcus pneumoniae*: crystal structure and evaluation of plasmin (ogen)-binding sites, *Journal of Molecular Biology*, **343**.4, 997–1005. 205, 206

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis, *Genome research*, **8**.3, 163–167. 107

Engelhardt, B. E., Jordan, M. I. and Brenner, S. E., (2006). A graphical model for predicting protein molecular function. In *Proceedings of the 23rd international conference on Machine learning*, pages 297–304. ACM. 108, 117

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Research*, **30**.7, 1575–1584. 65

Espadaler, J., Eswar, N., Querol, E., Avilés, F. X., Sali, A., Marti-Renom, M. A. and Oliva, B. (2008). Prediction of enzyme function by combining sequence similarity and protein interactions, *BMC Bioinformatics*, **9**.1, 249. 202

Falda, M., Toppo, S., Pescarolo, A., Lavezzo, E., Di Camillo, B., Facchinetti, A., Cilia, E., Velasco, R. and Fontana, P. (2012). Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms, *BMC Bioinformatics*, **13**.suppl 4, S14. 117

Fang, H. and Gough, J. (2013). A domain-centric solution to functional genomics via dcGO Predictor, *BMC Bioinformatics*, **14**.suppl 3, S9. 105, 106, 118, 234

Fayech, S., Essoussi, N. and Limam, M. (2009). Partitioning clustering algorithms for protein sequence data sets, *BioData mining*, **2**.1, 1–11. 63

Fenollar-Ferrer, C., Frau, J., Donoso, J. and Muñoz, F. (2008). Evolution of class C β-lactamases: factors influencing their hydrolysis and recognition mechanisms, *Theoretical Chemistry Accounts*, **121**.3-4, 209–218. 175

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J. et al. (2014). Pfam: the protein families database, *Nucleic Acids Research*, **42**.D1, D222–D230. 44, 66, 67, 104, 125

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A. et al. (2015). The Pfam protein families database: towards a more sustainable future, *Nucleic Acids Research*, **44**.D1, D279–D285. 46, 47

Fisher, A. B. (2011). Peroxiredoxin 6: a bifunctional enzyme with glutathione peroxidase and phospholipase a2 activities, *Antioxidants & redox signaling*, **15**.3, 831–844. 206

Forslund, K. and Sonnhammer, E. L. (2008). Predicting protein function from domain content, *Bioinformatics*, **24**.15, 1681–1687. 105

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points, *science*, **315**.5814, 972–976. 64

Friedberg, I. and Radivojac, P. (2016). Community-Wide Evaluation of Computational Function Prediction, *arXiv preprint arXiv:1601.01048*, **1601.01048**. 113, 115, 150

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics*, **28**.23, 3150–3152. 65, 71

Furnham, N., Dawson, N. L., Rahman, S. A., Thornton, J. M. and Orengo, C. A. (2015). Large-Scale Analysis Exploring Evolution of Catalytic Machineries and Mechanisms in Enzyme Superfamilies, *Journal of Molecular Biology*, **428**.2, 253–267. 55, 153, 164

Gallo Cassarino, T., Bordoli, L. and Schwede, T. (2014). Assessment of ligand binding site predictions in CASP10, *Proteins: Structure, Function, and Bioinformatics*, **82**.S2, 154–163. 157, 192

Galperin, M. Y. and Koonin, E. V. (2012). Divergence and convergence in enzyme evolution, *Journal of Biological Chemistry*, **287**.1, 21–28. 52

Galperin, M. Y., Makarova, K. S., Wolf, Y. I. and Koonin, E. V. (2014). Expanded microbial genome coverage and improved protein family annotation in the COG database, *Nucleic Acids Research*, **43**.D1, D261–D269. 47

Gancedo, C. and Flores, C.-L. (2008). Moonlighting proteins in yeasts, *Microbiology and Molecular Biology Reviews*, **72**.1, 197–210. 210

Garcia, A. M., Dawson, N. L., Kruger, F. A., Overington, J., Orengo, C. and Ranea, J. A. G. (2016). A Structural and Functional View of Polypharmacology, *bioRxiv*, page 044289. 100, 182, 234

Gerlt, J. A., Babbitt, P. C. and Rayment, I. (2005). Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity, *Archives of biochemistry and biophysics*, **433**.1, 59–70. 153

Glaser, F., Morris, R. J., Najmanovich, R. J., Laskowski, R. A. and Thornton, J. M. (2006). A method for localizing ligand binding pockets in protein structures, *PROTEINS: Structure, Function, and Bioinformatics*, **62**.2, 479–488. 156

Glaser, F., Rosenberg, Y., Kessel, A., Pupko, T. and Ben-Tal, N. (2005). The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures, *PROTEINS: Structure, Function, and Bioinformatics*, **58**.3, 610–617. 156

Godzik, A., Jambon, M. and Friedberg, I. (2007). Computational protein function prediction: Are we making progress?, *Cellular and Molecular Life Sciences*, **64**.19-20, 2505–2511. 110, 111

Gómez, A., Domedel, N., Cedano, J., Piñol, J. and Querol, E. (2003). Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins?, *Bioinformatics*, **19**.7, 895–896. 201, 202, 221, 225

Gómez, A., Hernández, S., Amela, I., Piñol, J., Cedano, J. and Querol, E. (2011). Do protein–protein interaction databases identify moonlighting proteins?, *Molecular BioSystems*, **7**.8, 2379–2382. 202

Gong, Q., Ning, W. and Tian, W. (2016). GoFDR: A sequence alignment based method for predicting protein functions, *Methods*, **93**, 3–14. 107

Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins, *Proceedings of the National Academy of Sciences*, **84**.13, 4355–4358. 42

Guralnik, V. and Karypis, G., (2001). A scalable algorithm for clustering sequential data. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 179–186. IEEE. 64

Haeggström, J. Z. (2004). Leukotriene A4 hydrolase/aminopeptidase, the gatekeeper of chemotactic leukotriene B4 biosynthesis, *Journal of Biological Chemistry*, **279**.49, 50639–50642. 213, 215

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K. and Beck, E. (2013). TIGRFAMs and genome properties in 2013, *Nucleic Acids Research*, **41**.D1, D387–D395. 46

Hall, B. G. and Barlow, M. (2004). Evolution of the serine β-lactamases: past, present and future, *Drug Resistance Updates*, **7**.2, 111–123. 174

Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A. and Clarke, J. (2007). The folding and evolution of multidomain proteins, *Nature Reviews Molecular Cell Biology*, **8**.4, 319–330. 25

Hannenhalli, S. S. and Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments, *Journal of Molecular Biology*, **303**.1, 61–76. 40

Hauser, M., Mayer, C. E. and Söding, J. (2013). kClust: fast and sensitive clustering of large protein sequence databases, *BMC Bioinformatics*, **14**.1, 248. 65

Hawkins, T., Chitale, M., Luban, S. and Kihara, D. (2009). PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data, *Proteins: Structure, Function, and Bioinformatics*, **74**.3, 566–582. 106, 202

Hawkins, T., Luban, S. and Kihara, D. (2006). Enhanced automated function prediction using distantly related sequences and contextual association by PFP, *Protein Science*, **15**.6, 1550–1556. 106, 107, 203

Hayete, B. and Bienkowska, J. R., (2005). Gotrees: predicting go associations from protein domain composition using decision trees. In *Pacific Symposium on Biocomputing*, volume 10, pages 127–138. World Scientific. 105

Heger, A. and Holm, L. (2003). Exhaustive enumeration of protein domain families, *Journal of Molecular Biology*, **328**.3, 749–767. 66

Hegyi, H. and Gerstein, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins, *Genome research*, **11**.10, 1632–1640. 103

Henderson, B., Fares, M. A. and Lund, P. A. (2013). Chaperonin 60: a paradoxical, evolutionarily conserved protein family with multiple moonlighting functions, *Biological Reviews*, **88**.4, 955–987. 223

Hendlich, M., Rippmann, F. and Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins, *Journal of Molecular Graphics and Modelling*, **15**.6, 359–363. 156

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks, *Proceedings of the National Academy of Sciences*, **89**.22, 10915–10919. 33, 66

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities, *Biochemical Journal*, **280**.2, 309–316. 68

Hernández, S., Ferragut, G., Amela, I., Perez-Pons, J., Piñol, J., Mozo-Villarias, A., Cedano, J. and Querol, E. (2014). MultitaskProtDB: a database of multitasking proteins, *Nucleic Acids Research*, **42**.D1, D517–D520. 201, 221

Hernández, S., Franco, L., Calvo, A., Ferragut, G., Hermoso, A., Amela, I., Gómez, A., Querol, E. and Cedano, J. (2015). Bioinformatics and moonlighting proteins, *Frontiers in bioengineering and biotechnology*, **3**. 201, 202, 221, 223

Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992). Selection of representative protein data sets, *Protein Science*, **1**.3, 409–417. 65

Holliday, G. L., Almonacid, D. E., Bartlett, G. J., O'Boyle, N. M., Torrance, J. W., Murray-Rust, P., Mitchell, J. B. and Thornton, J. M. (2007). MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms, *Nucleic Acids Research*, **35**.suppl 1, D515–D520. 26, 175

Holm, L. and Sander, C. (1994). Parser for protein folding units, *Proteins: Structure, Function, and Bioinformatics*, **19**.3, 256–268. 48

Holm, L. and Sander, C. (1995). Dali: a network tool for protein structure comparison, *Trends in biochemical sciences*, **20**.11, 478–480. 44, 108

Hu, S., Xie, Z., Onishi, A., Yu, X., Jiang, L., Lin, J., Rho, H.-s., Woodard, C., Wang, H., Jeong, J.-S. et al. (2009). Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling, *Cell*, **139**.3, 610–622. 206, 208

Huang, B. and Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation, *BMC Structural Biology*, **6**.1, 1–11. 156, 157

Huang, H., Pandya, C., Liu, C., Al-Obaidi, N. F., Wang, M., Zheng, L., Keating, S. T., Aono, M., Love, J. D., Evans, B. et al. (2015). Panoramic view of a superfamily of phosphatases through substrate profiling, *Proceedings of the National Academy of Sciences*, **112**.16, E1974–E1983. 54

Hubbard, S. J. and Thornton, J. M. (1993). Naccess, *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, **2**.1. 185

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M. et al. (2015). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences, *Nucleic Acids Research*, **44**.D1, D286–D293. 47

Innis, C. A., Anand, A. P. and Sowdhamini, R. (2004). Prediction of functional sites in proteins using conserved functional group analysis, *Journal of Molecular Biology*, **337**.4, 1053–1068. 151

Jeffery, C. J. (1999). Moonlighting proteins, *Trends in biochemical sciences*, **24**.1, 8–11. 59, 199, 200

Jeffery, C. J. (2003). Multifunctional proteins: examples of gene sharing, *Annals of medicine*, **35**.1, 28–35. 109, 199, 200

Jeffery, C. J. (2004a). Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins, *Current Opinion in Structural Biology*, **14**.6, 663–668. 201

Jeffery, C. J. (2004b). Moonlighting proteins: complications and implications for proteomics research, *Drug Discovery Today: Targets*, **3**.2, 71–78. 200, 206

Jeffery, C. J. (2011). Proteins with neomorphic moonlighting functions in disease, *IUBMB life*, **63**.7, 489–494. 211

Jeffery, C. J. (2015). Why study moonlighting proteins?, *Frontiers in genetics*, **6**. 226

Jeffery, C. J., Bahnson, B. J., Chien, W., Ringe, D. and Petsko, G. A. (2000). Crystal structure of rabbit phosphoglucose isomerase, a glycolytic enzyme that moonlights as neuroleukin, autocrine motility factor, and differentiation mediator, *Biochemistry*, **39**.5, 955–964. 217, 218

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy, *arXiv preprint*, **cmp-lg/9709008**. 28

Jiang, Y., Clark, W. T., Friedberg, I. and Radivojac, P. (2014). The impact of incomplete knowledge on the evaluation of protein function prediction: a structured-output learning perspective, *Bioinformatics*, **30**.17, i609–i616. 118, 119, 120

Jiang, Y., Oron, R. T., Clark, T. W., Bankapur, R. A., D'Andrea, D., Lepore, R., Funk, S. C., Kahanda, I., Verspoor, M. K., Ben-Hur, A., Koo, E. D. C., Penfold-Brown, D., Shasha, D., Youngs, N., Bonneau, R., Lin, A., Sahraeian, E. S. M., Martelli, L. P., Profiti, G., Casadio, R., Cao, R., Zhong, Z., Cheng, J., Altenhoff, A., Skunca, N., Dessimoz, C., Dogan, T., Hakala, K., Kaewphan, S., Mehryary, F., Salakoski, T., Ginter, F., Fang, H., Smithers, B., Oates, M., Gough, J., Törönen, P., Koskinen, P., Holm, L., Chen, C.-T., Hsu, W.-L., Bryson, K., Cozzetto, D., Minneci, F., Jones, T. D., Chapman, S., BKC, D., Khan, K. I., Kihara, D., Ofer, D., Rappoport, N., Stern, A., Cibrian-Uhalte, E., Denny, P., Foulger, E. R., Hieta, R., Legge, D., Lovering, C. R., Magrane, M., Melidoni, N. A., Mutowo-Meullenet, P., Pichler, K., Shypitsyna, A., Li, B., Zakeri, P., ElShal, S., Tranchevent, L.-C., Das, S., Dawson, L. N., Lee, D., Lees, G. J., Sillitoe, I., Bhat, P., Nepusz, T., Romero, E. A., Sasidharan, R., Yang, H., Paccanaro, A., Gillis, J., Sedeño-Cortés, E. A., Pavlidis, P., Feng, S., Cejuela, M. J., Goldberg, T., Hamp, T., Richter, L., Salamov, A., Gabaldon, T., Marcet-Houben, M., Supek, F., Gong, Q., Ning, W., Zhou, Y., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Toppo, S., Ferrari, C., Giollo, M., Piovesan, D., Tosatto, C. S., del Pozo, A., Fernández, M. J., Maietta, P., Valencia, A., Tress, L. M., Benso, A., Di Carlo, S., Politano, G., Savino, A., Rehman, U. H., Re, M., Mesiti, M., Valentini, G., Bargsten, W. J., van Dijk, J. A. D., Gemovic, B., Glisic, S., Perovic, V., Veljkovic, V., Veljkovic, N., Almeida-e Silva, C. D., Vencio, N. R. Z., Sharan, M., Vogel, J., Kansakar, L., Zhang, S., Vucetic, S., Wang, Z., Sternberg, E. M. J., Wass, N. M., Huntley, P. R., Martin, J. M., O'Donovan, C., Robinson, N. P., Moreau, Y., Tramontano, A., Babbitt, C. P., Brenner, E. S., Linial, M., Orengo, A. C., Rost, B., Greene, S. C., Mooney, D. S., Friedberg, I. and Radivojac, P. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy, *Genome Biology*, **17**.1, 1–19. 113, 133, 134, 135, 142, 234

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences, *Computer applications in the biosciences: CABIOS*, **8**.3, 275–282. 34, 37

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G. et al. (2014). InterProScan 5: genome-scale protein function classification, *Bioinformatics*, **30**.9, 1236–1240. 151

Jones, S. and Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches, *Journal of Molecular Biology*, **272**.1, 121–132. 154

Jores, R., Alzari, P. M. and Meo, T. (1990). Resolution of hypervariable regions in t-cell receptor beta chains by a modified wu-kabat index of amino acid diversity., *Proceedings of the National Academy of Sciences*, **87**.23, 9138–9142. 37

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation, *Nucleic Acids Research*, **44**.D1, D457–D462. 25

Kannan, N., Taylor, S. S., Zhai, Y., Venter, J. C. and Manning, G. (2007). Structural and functional diversity of the microbial kinome, *PLoS Biology*, **5**.3, e17. 56

Karlin, S. and Brocchieri, L. (1996). Evolutionary conservation of reca genes in relation to protein structure and function., *Journal of bacteriology*, **178**.7, 1881–1894. 37

Katoh, K., Misawa, K., Kuma, K.-i. and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Research*, **30**.14, 3059–3066. 36, 71, 90, 187

Kerfeld, C. A. and Scott, K. M. (2011). Using BLAST to teach "E-value-tionary" concepts, *PLoS Biology*, **9**.2, e1001014. 41

Khan, I. K., Chitale, M., Rayon, C. and Kihara, D., (2012). Evaluation of function predictions by PFP, ESG, and PSI-BLAST for moonlighting proteins. In *BMC proceedings*, volume 6, page S5. BioMed Central Ltd. 201, 202, 221, 223, 225

Khan, I. K., Wei, Q., Chitale, M. and Kihara, D. (2015). Pfp/esg: automated protein function prediction servers enhanced with gene ontology visualization tool, *Bioinformatics*, **31**.2, 271–272. 203

Kinhikar, A. G., Vargas, D., Li, H., Mahaffey, S. B., Hinds, L., Belisle, J. T. and Laal, S. (2006). Mycobacterium tuberculosis malate synthase is a laminin-binding adhesin, *Molecular microbiology*, **60**.4, 999–1013. 210

Knauer, S. H., Artsimovitch, I. and Rösch, P. (2012). Transformer proteins, *Cell Cycle*, **11**.23, 4289–4290. 220

Kohl, M., Wiese, S. and Warscheid, B., (2011). Cytoscape: software for visualization and analysis of biological networks. In *Data Mining in Proteomics*, pages 291–303. Springer. 171, 172

Koskinen, P., Törönen, P., Nokso-Koivisto, J. and Holm, L. (2015). PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment, *Bioinformatics*, **31**.10, 1544–1552. 117

Krause, A., Stoye, J. and Vingron, M. (2005). Large scale hierarchical clustering of protein sequences, *BMC Bioinformatics*, **6**.1, 15. 63

Krishnamurthy, N., Brown, D. P., Kirshner, D. and Sjölander, K. (2006). PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification, *Genome biology*, **7**.9, R83. 66

Krogh'f, A. and Brown, I. (1994). Hidden Markov Models in Computational Biology, *J. Mol. Bioi*, **235**, 1501–1531. 42

Kruskal, W. H. (1957). Historical notes on the wilcoxon unpaired two-sample test, *Journal of the American Statistical Association*, **52**.279, 356–360. 98, 189, 190

Lam, S. D., Dawson, N. L., Das, S., Sillitoe, I., Ashford, P., Lee, D., Lehtinen, S., Orengo, C. A. and Lees, J. G. (2016). Gene3D: expanding the utility of domain assignments, *Nucleic Acids Research*, **44**.D1, D404–D409. 234

Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions, *Journal of molecular graphics*, **13**.5, 323–330. 156, 183, 186

Laskowski, R. A., Watson, J. D. and Thornton, J. M. (2003). From protein structure to biochemical function?, *Journal of structural and functional genomics*, **4**.2-3, 167–177. 109

Lee, D., Das, S., Dawson, N. L., Dobrijevic, D., Ward, J. and Orengo, C. (2016). Novel computational protocols for functionally classifying and characterising serine beta-lactamases, *PLoS Comput Biol*, **12**.6, e1004926. 179

Lee, D. A., Rentzsch, R. and Orengo, C. (2010). GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains, *Nucleic Acids Research*, **38**.3, 720–737. 67, 68, 70, 71, 72, 73, 77, 93, 196, 227, 232

Lees, J. G., Lee, D., Studer, R. A., Dawson, N. L., Sillitoe, I., Das, S., Yeats, C., Dessailly, B. H., Rentzsch, R. and Orengo, C. A. (2014). Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis, *Nucleic Acids Research*, **42**.D1, D240–D245. 50, 70, 76

Letunic, I., Doerks, T. and Bork, P. (2015). SMART: recent updates, new developments and status in 2015, *Nucleic Acids Research*, **43**.D1, D257–D260. 46

Levitt, M. (2009). Nature of the protein universe, *Proceedings of the National Academy of Sciences*, **106**.27, 11079–11084. 47

Lichtarge, O., Bourne, H. R. and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families, *Journal of Molecular Biology*, **257**.2, 342–358. 40, 151, 154, 155, 192

Lim, M. L., Lum, M.-G., Hansen, T. M., Roucou, X. and Nagley, P. (2002). On the release of cytochromec from mitochondria during cell death signaling, *Journal of biomedical science*, **9**.6, 488–506. 206

Lin, D., (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, volume 2, pages 768–774. Association for Computational Linguistics. 28

Liu, J. and Rost, B. (2003). Domains, motifs and clusters in the protein universe, *Current Opinion in Chemical Biology*, **7**.1, 5–11. 63

Lopez, D. and Pazos, F. (2013). Concomitant prediction of function and fold at the domain level with GO-based profiles, *BMC Bioinformatics*, **14**.suppl 3, S12. 105

Lua, R. C., Wilson, S. J., Konecki, D. M., Wilkins, A. D., Venner, E., Morgan, D. H. and Lichtarge, O. (2016). UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures, *Nucleic Acids Research*, **44**.D1, D308–D312. 192

Luo, X., Hsiao, H.-H., Bubunenko, M., Weber, G., Court, D. L., Gottesman, M. E., Urlaub, H. and Wahl, M. C. (2008). Structural and functional analysis of the *E. coli* NusB-S10 transcription antitermination complex, *Molecular Cell*, **32**.6, 791–802. 217

Madera, M. (2008). Profile Comparer: a program for scoring and aligning profile hidden Markov models, *Bioinformatics*, **24**.22, 2630–2631. 170, 171

Mani, M., Chen, C., Amblee, V., Liu, H., Mathur, T., Zwicke, G., Zabad, S., Patel, B., Thakkar, J. and Jeffery, C. J. (2014). MoonProt: a database for proteins that are known to moonlight, *Nucleic Acids Research*, page gku954. 201, 204

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I. et al. (2014). CDD: NCBI's conserved domain database, *Nucleic Acids Research*, **43**.D1, D222–D226. 46, 105, 125

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B., Taroni, C. and Thornton, J. M. (1998). Protein folds and functions, *Structure*, **6**.7, 875–884. 45, 108

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta (BBA)-Protein Structure*, **405**.2, 442–451. 158

Mazin, P. V., Gelfand, M. S., Mironov, A. A., Rakhmaninova, A. B., Rubinov, A. R., Russell, R. B. and Kalinina, O. V. (2010). An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies., *Algorithms for Molecular Biology*, **5**.1, 29. 77

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. and Thomas, P. D. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools, *Nucleic Acids Research*, **44**.D1, D336–D342. 46

Mistry, M. and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity, *BMC bioinformatics*, **9**.1, 1. 29

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S. et al. (2014). The InterPro protein families database: the classification resource after 15 years, *Nucleic Acids Research*, **43**.D1, D213–221. 46

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, **247**.4, 536–540. 23, 44, 50, 108

Nagao, C., Nagano, N. and Mizuguchi, K. (2010). Relationships between functional subclasses and information contained in active-site and ligand-binding residues in diverse superfamilies, *Proteins: Structure, Function, and Bioinformatics*, **78**.10, 2369–2384. 152

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology*, **48**.3, 443–453. 34, 35

Nikolaev, Y., Deillon, C., Hoffmann, S. R., Bigler, L., Friess, S., Zenobi, R., Pervushin, K., Hunziker, P. and Gutte, B. (2010). The leucine zipper domains of the transcription factors GCN4 and c-Jun have ribonuclease activity, *PloS One*, **5**.5, e10765. 206

Notredame, C., Higgins, D. G. and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment, *Journal of Molecular Biology*, **302**.1, 205–217. 36

Nuin, P. A., Wang, Z. and Tillier, E. R. (2006). The accuracy of several multiple sequence alignment programs for proteins, *BMC Bioinformatics*, **7**.1, 471. 36

Ojha, S., Meng, E. C. and Babbitt, P. C. (2007). Evolution of function in the "two dinucleotide binding domains" flavoproteins, *PLoS Computational Biology*, **3**.7, e121. 56

Ondrechen, M. J., Clifton, J. G. and Ringe, D. (2001). THEMATICS: a simple computational predictor of enzyme function from structure, *Proceedings of the National Academy of Sciences*, **98**.22, 12473–12478. 108

Orengo, C. A., Michie, A., Jones, S., Jones, D. T., Swindells, M. and Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures, *Structure*, **5**.8, 1093–1109. 23, 48, 49, 108

Orengo, C. A. and Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison, *Methods in Enzymology*, **266**, 617–635. 45

Paccanaro, A., Casbon, J. A. and Saqi, M. A. (2006). Spectral clustering of protein sequences, *Nucleic Acids Research*, **34**.5, 1571–1580. 64

Pal, D. and Eisenberg, D. (2005). Inference of protein function from protein structure, *Structure*, **13**.1, 121–130. 109

Pandya, C., Farelli, J. D., Dunaway-Mariano, D. and Allen, K. N. (2014). Enzyme promiscuity: engine of evolutionary innovation, *Journal of Biological Chemistry*, **289**.44, 30229–30236. 59

Pearson, W. R., (1994). Using the FASTA program to search protein and DNA sequence databases. In *Computer Analysis of Sequence Data*, pages 307–331. Springer. 41

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., de Castro, E., Baratin, D., Cuche, B. A., Bougueleret, L., Poux, S. et al. (2015). HAMAP in 2015: updates to the protein family classification and annotation system, *Nucleic Acids Research*, **43**.D1, D1064–D1070. 46

Pethica, R. B., Levitt, M. and Gough, J. (2012). Evolutionarily consistent families in SCOP: sequence, structure and function, *BMC Structural Biology*, **12**.1, 27. 50

Petryszak, R., Kretschmann, E., Wieser, D. and Apweiler, R. (2005). The predictive power of the CluSTr database, *Bioinformatics*, **21**.18, 3604–3609. 63

Piovesan, D., Martelli, P. L., Fariselli, P., Profiti, G., Zauli, A., Rossi, I. and Casadio, R. (2013). How to inherit statistically validated annotation within BAR+ protein clusters, *BMC Bioinformatics*, **14**.suppl 3, S4. 104, 117

Piovesan, D., Martelli, P. L., Fariselli, P., Zauli, A., Rossi, I. and Casadio, R. (2011). BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences, *Nucleic Acids Research*, **39**.suppl 2, W197–W202. 104

Porter, C. T., Bartlett, G. J. and Thornton, J. M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Research*, **32**.suppl 1, D129–D133. 26, 44, 92, 98, 164, 186, 204

Prieto, C. and De Las Rivas, J. (2006). APID: agile protein interaction dataanalyzer, *Nucleic Acids Research*, **34**.suppl 2, W298–W302. 202

Punta, M. and Ofran, Y. (2008). The rough guide to *in silico* function prediction, or how to use sequence and structure information to predict protein function, *PLoS Computational Biology*, **4**.10, e1000160. 109

R-Core-Team (2014). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012, *Open access available at: http://cran. r-project. org*. 98, 189, 190

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A. et al. (2013). A large-scale

evaluation of computational protein function prediction, *Nature Methods*, **10**.3, 221–227. 113, 117, 118, 131, 234

Rappoport, N., Linial, N. and Linial, M. (2013). ProtoNet: charting the expanding universe of protein sequences, *Nature Biotechnology*, **31**.4, 290–292. 63, 64

Rappoport, N., Stern, A., Linial, N. and Linial, M. (2014). Entropy-driven partitioning of the hierarchical protein space, *Bioinformatics*, **30**.17, i624–i630. 64

Rausell, A., Juan, D., Pazos, F. and Valencia, A. (2010). Protein interactions and ligand binding: from protein subfamilies to functional specificity, *Proceedings of the National Academy of Sciences*, **107**.5, 1995–2000. 39, 77

Read, J., Pearce, J., Li, X., Muirhead, H., Chirgwin, J. and Davies, C. (2001). The crystal structure of human phosphoglucose isomerase at 1.6 å resolution: implications for catalytic mechanism, cytokine activity and haemolytic anaemia, *Journal of Molecular Biology*, **309**.2, 447–463. 217, 218

Reddy, T., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E. A. and Kyrpides, N. C. (2014). The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta) genome project classification, *Nucleic Acids Research*, **43**.D1, D1099–D1106. 102

Redfern, O. C., Harrison, A., Dallman, T., Pearl, F. M. and Orengo, C. A. (2007). CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures, *PLoS Computational Biology*, **3**.11, e232. 48, 49

Reeves, G. A., Dallman, T. J., Redfern, O. C., Akpor, A. and Orengo, C. A. (2006). Structural diversity of domain superfamilies in the CATH database, *Journal of Molecular Biology*, **360**.3, 725–741. 52, 53

Reid, A. J., Ranea, J. A. and Orengo, C. A. (2010). Comparative evolutionary analysis of protein complexes in *E. coli* and yeast, *BMC Genomics*, **11**.1, 1–16. 52

Rentzsch, R., (2012). *Protocols to capture the functional plasticity of protein domain superfamilies*. PhD thesis, University College London. 72, 232, 233

Rentzsch, R. and Orengo, C. A. (2009). Protein function prediction–the power of multiplicity, *Trends in biotechnology*, **27**.4, 210–219. 28

Rentzsch, R. and Orengo, C. A. (2013). Protein function prediction using domain families, *BMC Bioinformatics*, **14**.suppl 3, S5. 74, 105, 106, 227, 234

Resnik, P. and Yarowsky, D. (1999). Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation, *Natural language engineering*, **5**.02, 113–133. 28

Reva, B., Antipin, Y. and Sander, C. (2007). Determinants of protein function revealed by combinatorial entropy optimization, *Genome biology*, **8**.11, 1. 40

Richmond, T. J. and Richards, F. M. (1978). Packing of α-helices: Geometrical constraints and contact areas, *Journal of molecular biology*, **119**.4, 537–555. 185

Roche, D. B., Tetchner, S. J. and McGuffin, L. J. (2010). The binding site distance test score: a robust method for the assessment of predicted protein binding sites, *Bioinformatics*, **26**.22, 2920–2921. 158

Rojas, A. M., Fuentes, G., Rausell, A. and Valencia, A. (2012). The Ras protein superfamily: evolutionary tree and role of conserved amino acids, *The Journal of Cell Biology*, **196**.2, 189–201. 39

Rose, P. W., Prlić, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., Green, R. K., Goodsell, D. S., Westbrook, J. D., Woo, J., Young, J., Zardecki, C., Berman, H. M., Bourne, P. E. and Burley, S. K. (2015). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education, *Nucleic Acids Research*, **43**.D1, D345–D356. 30

Rost, B. (2002). Enzyme function less conserved than anticipated, *Journal of Molecular Biology*, **318**.2, 595–608. 103

Royle, S. J. (2013). Protein adaptation: mitotic functions for membrane trafficking proteins, *Nature Reviews Molecular Cell Biology*, **14**.9, 592–599. 226

Rudberg, P. C., Tholander, F., Thunnissen, M. M. and Haeggström, J. Z. (2002). Leukotriene A4Hydrolase/Aminopeptidase Glutamate 271 is a catalytic residue with specific roles in two distinct enzyme mechanisms, *Journal of Biological Chemistry*, **277**.2, 1398–1404. 215

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M. et al. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids Research*, **32**.18, 5539–5545. 25

Sadreyev, R. and Grishin, N. (2003). Compass: a tool for comparison of multiple protein alignments with assessment of statistical significance, *Journal of Molecular Biology*, **326**.1, 317–336. 43, 71

Sandhya, S., Rani, S. S., Pankaj, B., Govind, M. K., Offmann, B., Srinivasan, N. and Sowdhamini, R. (2009). Length variations amongst protein domain superfamilies and consequences on structure and function, *PLoS One*, **4**.3, No–PP. 53

Sankararaman, S. and Sjölander, K. (2008). INTREPID—INformation-theoretic TREe traversal for Protein functional site IDentification, *Bioinformatics*, **24**.21, 2445–2452. 154

Scaiewicz, A. and Levitt, M. (2015). The language of the protein universe, *Current Opinion in Genetics & Development*, **35**, 50–56. 47

Scheerer, P., Borchert, A., Krauss, N., Wessner, H., Gerth, C., Höhne, W. and Kuhn, H. (2007). Structural Basis for Catalytic Activity and Enzyme Polymerization of Phospholipid Hydroperoxide Glutathione Peroxidase-4 (GPx4), *Biochemistry*, **46**.31, 9041–9049. 217

Schlicker, A., Domingues, F. S., Rahnenführer, J. and Lengauer, T. (2006). A new measure for functional similarity of gene products based on gene ontology, *BMC Bioinformatics*, **7**.1, 302. 28

Schmidt, T., Haas, J., Cassarino, T. G. and Schwede, T. (2011). Assessment of ligand-binding residue predictions in CASP9, *Proteins: Structure, Function, and Bioinformatics*, **79**.S10, 126–136. 157, 192

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C. and Friedberg, I. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space, *PLoS Computational Biology*, **9**.5, e1003063. 29, 30, 119

Schug, J., Diskin, S., Mazzarelli, J., Brunk, B. P. and Stoeckert, C. J. (2002). Predicting gene ontology functions from ProDom and CDD protein domains, *Genome research*, **12**.4, 648–655. 105

Shannon, C. E. (2001). A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**.1, 3–55. 37

Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path., *Protein Engineering*, **11**.9, 739–747. 44, 108

Shoemaker, B. A., Zhang, D., Tyagi, M., Thangudu, R. R., Fong, J. H., Marchler-Bauer, A., Bryant, S. H., Madej, T. and Panchenko, A. R. (2012). IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins, *Nucleic Acids Research*, **40**.D1, D834–D840. 187

Shulman-Peleg, A., Nussinov, R. and Wolfson, H. J. (2005). SiteEngines: recognition and comparison of binding sites and protein–protein interfaces, *Nucleic Acids Research*, **33**.suppl 2, W337–W341. 108

Siddiqui, A. S. and Barton, G. J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions, *Protein Science*, **4**.5, 872–884. 48

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J. et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Molecular Systems Biology*, **7**.1. 36

Sigrist, C. J., De Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L. and Xenarios, I. (2012). New and continuing developments at PROSITE, *Nucleic Acids Research*, **41**.D1, D344–D347. 46

Sillitoe, I., Cuff, A. L., Dessailly, B. H., Dawson, N. L., Furnham, N., Lee, D., Lees, J. G., Lewis, T. E., Studer, R. A., Rentzsch, R. et al. (2013). New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures, *Nucleic Acids Research*, **41**.D1, D490–D498. 100

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G. et al. (2015). CATH: comprehensive structural and functional annotations for genome sequences, *Nucleic Acids Research*, **43**.D1, D376–D381. 44, 49, 90, 104, 123

Sjolander, K., (1998). Phylogenetic inference in protein superfamilies: analysis of SH2 domains. In *Proc. Int. Conf. Intell. Syst. Mol. Biol*, volume 6, pages 165–174. 66

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S. and Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology, *Computer applications in the biosciences: CABIOS*, **12**.4, 327–345. 66

Škunca, N., Altenhoff, A. and Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations, *PLoS Computational Biology*, **8**.5, e1002533. 74, 161

Smith, M., Kunin, V., Goldovsky, L., Enright, A. J. and Ouzounis, C. A. (2005). MagicMatch—cross-referencing sequence identifiers across databases, *Bioinformatics*, **21**.16, 3429–3430. 125

Smith, T. F. and Waterman, M. S. (1981). Comparison of biosequences, *Advances in Applied Mathematics*, **2**.4, 482–489. 34, 35

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization, *Bioinformatics*, **27**.3, 431–432. 170

Söding, J., Biegert, A. and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction, *Nucleic Acids Research*, **33**.suppl 2, W244–W248. 43

Sperisen, P. and Pagni, M. (2005). JACOP: a simple and robust method for the automated classification of protein sequences with modular architecture, *BMC Bioinformatics*, **6**.1, 1–12. 64

Steczkiewicz, K., Muszewska, A., Knizewski, L., Rychlewski, L. and Ginalski, K. (2012). Sequence, structure and functional diversity of PD-(D/E) XK phosphodiesterase superfamily, *Nucleic Acids Research*, **40**.15, 7016–7045. 59

Studer, R. A. and Robinson-Rechavi, M. (2009). How confident can we be that orthologs are similar, but paralogs differ?, *Trends in Genetics*, **25**.5, 210–216. 32

Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms, *PloS One*, **6**.7, e21800. 144

Swindells, M. B. (1995). A procedure for detecting structural domains in proteins, *Protein Science*, **4**.1, 103–112. 48

Takeuchi, R., Certo, M., Caprara, M. G., Scharenberg, A. M. and Stoddard, B. L. (2009). Optimization of in vivo activity of a bifunctional homing endonuclease and maturase reverses evolutionary degradation, *Nucleic Acids Research*, **37**.3, 877–890. 206

Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment, *Journal of Molecular Biology*, **208**.1, 1–22. 44, 48, 108, 177

Taylor, W. R. (1986). The classification of amino acid conservation, *Journal of theoretical Biology*, **119**.2, 205–218. 24, 37

Templeton, P. D., Litman, E. S., Metzner, S. I., Ahn, N. G. and Sousa, M. C. (2013). Structure of Mediator of RhoA-Dependent Invasion (MRDI) Explains Its Dual Function as a Metabolic Enzyme and a Mediator of Cell Invasion, *Biochemistry*, **52**.33, 5675–5684. 210, 212

Theißen, G. (2002). Orthology: secret life of genes, *Nature*, **415**.6873, 741–741. 109

Tholander, F., Muroya, A., Roques, B.-P., Fournié-Zaluski, M.-C., Thunnissen, M. M. and Haeggström, J. Z. (2008). Structure-based dissection of the active site chemistry of leukotriene A4 hydrolase: implications for M1 aminopeptidases and inhibitor design, *Chemistry & biology*, **15**.9, 920–929. 215

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., Blake, J. A., Consortium, G. O. et al. (2012). On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report, *PLoS Computational Biology*, **8**.2, e1002386. 119

Thompson, J. D., Gibson, T., Higgins, D. G. et al. (2002). Multiple sequence alignment using ClustalW and ClustalX, *Current Protocols in Bioinformatics*, pages 2–3. 36

Thompson, J. D., Linard, B., Lecompte, O. and Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives, *PloS One*, **6**.3, e18093. 36

Tian, W. and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity?, *Journal of molecular biology*, **333**.4, 863–882. 103

Tina, K., Bhadra, R. and Srinivasan, N. (2007). PIC: protein interactions calculator, *Nucleic Acids Research*, **35**.suppl 2, W473–W476. 186

Todd, A. E., Orengo, C. A. and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective, *Journal of Molecular Biology*, **307**.4, 1113–1143. 52, 55, 56, 70, 102, 103, 109

Tolbert, N. (1981). Metabolic pathways in peroxisomes and glyoxysomes, *Annual Review of Biochemistry*, **50**.1, 133–157. 209

Tuinstra, R. L., Peterson, F. C., Kutlesa, S., Elgin, E. S., Kron, M. A. and Volkman, B. F. (2008). Interconversion between two unrelated protein folds in the lymphotactin native state, *Proceedings of the National Academy of Sciences*, **105**.13, 5057–5062. 220

UniProt-Consortium (2015). UniProt: a hub for protein information, *Nucleic Acids Research*, **43**.D1, D204–D212. 30, 44, 50, 204, 233

Valdar, W. S. (2002). Scoring residue conservation, *Proteins: Structure, Function, and Bioinformatics*, **48**.2, 227–241. 24, 37, 82, 92, 93, 98, 146

Van Dongen, S. (2000). A cluster algorithm for graphs, *Report-Information systems*, .10, 1–40. 64

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. and Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins, *Current Opinion in Structural Biology*, **14**.2, 208–216. 23

Vogel, C. and Pleiss, J. (2014). The modular structure of ThDP-dependent enzymes, *Proteins: Structure, Function, and Bioinformatics*, **82**.10, 2523–2537. 56, 67

Wang, J. Z., Du, Z., Payattakool, R., Philip, S. Y. and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms, *Bioinformatics*, **23**.10, 1274–1281. 29

Ward, R. M., Venner, E., Daines, B., Murray, S., Erdin, S., Kristensen, D. M. and Lichtarge, O. (2009). Evolutionary Trace Annotation Server: automated enzyme function prediction in protein structures using 3D templates, *Bioinformatics*, **25**.11, 1426–1427. 117

Wass, M. N., Barton, G. and Sternberg, M. J. (2012). CombFunc: predicting protein function using heterogeneous data sources, *Nucleic Acids Research*, **40**.W1, W466–W470. 109, 129

Wass, M. N., David, A. and Sternberg, M. J. (2011). Challenges for the prediction of macromolecular interactions, *Current Opinion in Structural Biology*, **21**.3, 382–390. 153, 157

Wass, M. N., Kelley, L. A. and Sternberg, M. J. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures, *Nucleic Acids Research*, **38**.suppl 2, W469–W473. 157

Wass, M. N. and Sternberg, M. J. (2008). ConFunc—functional annotation in the twilight zone, *Bioinformatics*, **24**.6, 798–806. 107

Webb, E. C., (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Number Ed. 6. Academic Press. 26

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, *Molecular biology and evolution*, **18**.5, 691–699. 34

Wichelecki, D. J., Balthazor, B. M., Chau, A. C., Vetting, M. W., Fedorov, A. A., Fedorov, E. V., Lukk, T., Patskovsky, Y. V., Stead, M. B., Hillerich, B. S. et al. (2014). Discovery of function in the enolase superfamily: D-mannonate and D-gluconate dehydratases in the D-mannonate dehydratase subgroup, *Biochemistry*, **53**.16, 2722–2731. 55

Wicker, N., Perrin, G. R., Thierry, J. C. and Poch, O. (2001). Secator: a program for inferring protein subfamilies from phylogenetic trees, *Molecular Biology and Evolution*, **18**.8, 1435–1441. 66

Widmann, M., Radloff, R. and Pleiss, J. (2010). The Thiamine diphosphate dependent Enzyme Engineering Database: A tool for the systematic analysis of sequence and structure relations, *BMC biochemistry*, **11**.9, 1–6. 67, 76, 92

Wilkins, A., Erdin, S., Lua, R. and Lichtarge, O. (2012). Evolutionary trace for prediction and redesign of protein functional sites, *Computational Drug Discovery and Design*, pages 29–42. 155

Wilson, C. A., Kreychman, J. and Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores, *Journal of Molecular Biology*, **297**.1, 233–249. 102, 103

Wilson, K. P., Shewchuk, L. M., Brennan, R. G., Otsuka, A. J. and Matthews, B. W. (1992). *Escherichia coli* biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin-and DNA-binding domains., *Proceedings of the National Academy of Sciences*, **89**.19, 9257–9261. 210, 211, 212

Wu, T. T. and Kabat, E. A. (1970). An analysis of the sequences of the variable regions of bence jones proteins and myeloma light chains and their implications for antibody complementarity., *The Journal of experimental medicine*, **132**.2, 211–250. 37

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M. and Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Research*, **30**.1, 303–305. 202

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., Hönigschmid, P., Schafferhans, A., Roos, M., Bernhofer, M. et al. (2014). PredictProtein–an open resource for online prediction of protein structural and functional features, *Nucleic Acids Research*, **42**.W1, W337–W343. 109

Yeats, C., Redfern, O. C. and Orengo, C. (2010). A fast and automated solution for accurately resolving protein domain architectures, *Bioinformatics*, **26**.6, 745–751. 121, 125

Yona, G., Linial, N. and Linial, M. (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families, *Nucleic acids research*, **28**.1, 49–55. 63

Yun, M., Park, C.-G., Kim, J.-Y. and Park, H.-W. (2000). Structural analysis of glyceraldehyde 3-phosphate dehydrogenase from *Escherichia coli*: direct evidence of substrate binding and cofactor-induced conformational changes, *Biochemistry*, **39**.35, 10702–10710. 214

Zhao, B., Lei, L., Vassylyev, D. G., Lin, X., Cane, D. E., Kelly, S. L., Yuan, H., Lamb, D. C. and Waterman, M. R. (2009). Crystal structure of albaflavenone monooxygenase containing a moonlighting terpene synthase active site, *Journal of Biological Chemistry*, **284**.52, 36711–36719. 206, 207

Zhao, B., Lin, X., Lei, L., Lamb, D. C., Kelly, S. L., Waterman, M. R. and Cane, D. E. (2008). Biosynthesis of the sesquiterpene antibiotic albaflavenone in Streptomyces coelicolor A3 (2), *Journal of Biological Chemistry*, **283**.13, 8183–8189. 206, 207